Genetics and Genomics
Computational and Systems Biology

Massively Parallel Polyribosome Profiling Reveals Translation Defects of Human Disease-Relevant UTR Mutations

Reviewed Preprint
v2 • September 16, 2025
Revised by authors

Reviewed Preprint
v1 • August 27, 2024

Wei-Ping Li, Jia-Ying Su, Yu-Chi Chang, Hung-Lun Chiang, Yun-Lin Wang, Ang-Chu Huang, Yu-Tung Hsieh, Yi-Hsuan Chiang, Yen-Ling Ko, Bing-Jen Chiang, Cheng-Han Yang, Yen-Tsung Huang, Chien-Ling Lin

Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan • Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Vrije Universiteit Amsterdam, Amsterdam, Netherlands • Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

eLife Assessment

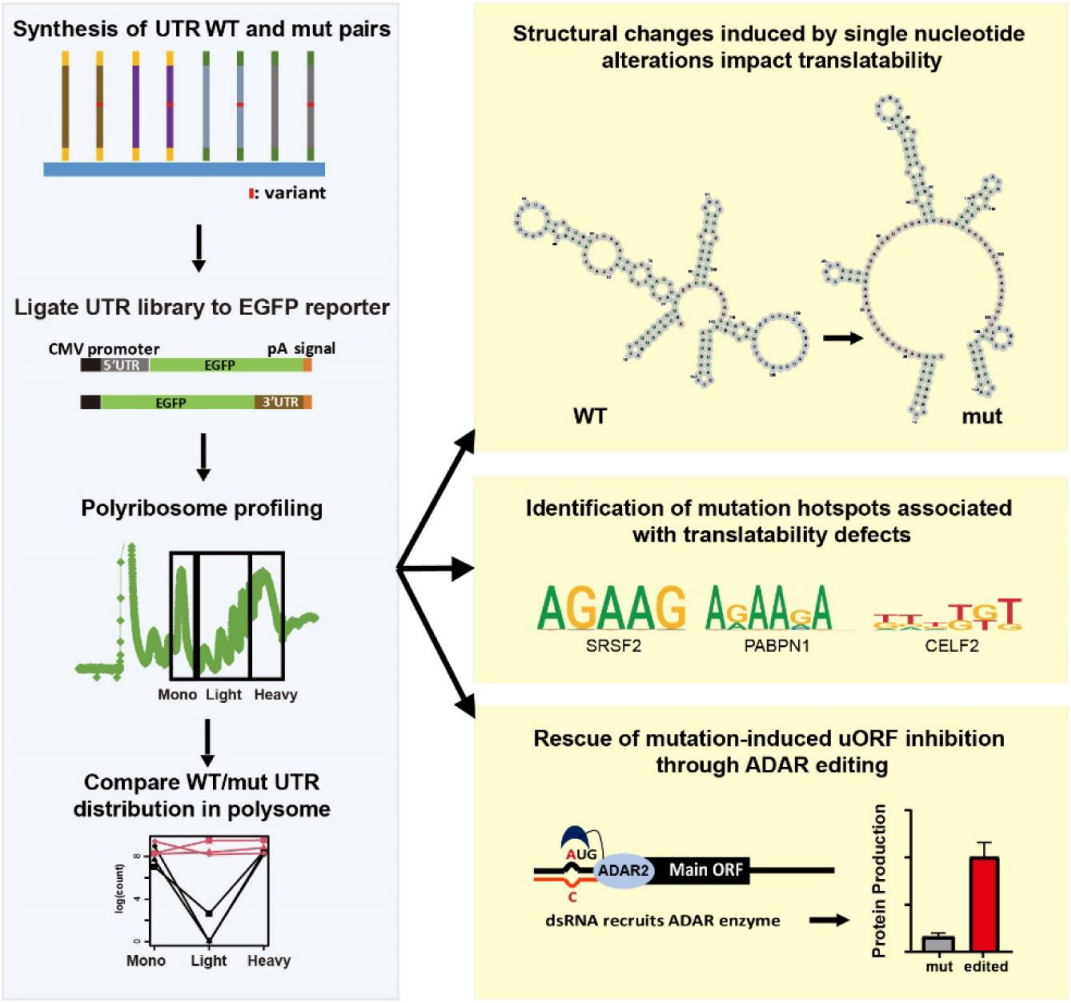
The effort is timely and the paper carries **valuable** insights into the function of UTR mutations. There are still significant concerns about both the quality of the screen data, and its ability to detect significant changes in translation and their direction. Therefore, the ability of the screen to support the extensive downstream statistical analysis is limited and leaves the paper **incomplete**.

https://doi.org/10.7554/eLife.98814.2.sa2

Abstract

The untranslated regions (UTRs) of mRNAs harbor regulatory elements influencing translation efficiency. Although 3.7% of disease-relevant human mutations occur in UTRs, their exact role in pathogenesis remains unclear. Through metagene analysis, we mapped pathogenic UTR mutations to regions near coding sequences, with a focus on the upstream open reading frame (uORF) initiation site. Subsequently, we utilized massively parallel poly(ribo)some profiling to compare the ribosome associations of 6,555 pairs of wildtype and mutant UTR fragments. We identified 46 UTR variants that altered polysome profiles, with enrichment in pathogenic mutations. Both univariate analysis and the elastic net regression model highlighted the significance of motifs of short repeated sequences, including SRSF2 binding sites, as mutation hotspots that lead to aberrant translation. Furthermore, these polysome-shifting mutations exhibited considerable impact on RNA secondary structures, particularly for upstream AUG-containing 5' UTRs. Integrating these features, our model achieved high accuracy (AUROC > 0.8) in predicting polysome-shifting mutations in the test dataset. Additionally, several lines of evidence indicate that changes in uORF usage underlie the translation deficiency arising from these mutations. Illustrating this, we demonstrate that a pathogenic mutation in the IRF6 5' UTR suppresses translation of the primary open reading frame by creating a uORF. Remarkably, site- directed ADAR editing of the mutant mRNA

rescued this translation deficiency. Overall, our study provides insights into the molecular mechanisms of UTR mutations and their links to clinical impacts through translation defects.



Introduction

Untranslated regions (UTRs) control translation efficiency

UTRs are crucial to gene expression regulation. Whilst mRNA coding regions determine protein sequence, the UTRs modulate protein abundance via, for instance, tuning stability and translation of mRNAs thereof. Many studies have shown that structural elements in 5' UTRs are involved in translation regulation. An iron regulatory protein (IRP) that maintains iron homeostasis binds to an iron-responsive element (IRE) in the 5' UTR of ferritin mRNAs, blocks recruitment of translation initiation factors, and elicits translation inhibition (Klausner et al., 1993; Muckenthaler et al., 1998). Internal ribosome entry sites (IRESs) are another *cis*-acting element of 5' UTRs that control translation. IRESs in 5' UTRs facilitate translation initiation by presenting ribosome-binding sites in a 5' cap-independent manner (Komar & Hatzoglou, 2005). On the contrary, guanine-quadruplex motifs, which can adopt a non-canonical secondary structure in the 5' UTR, are known to repress translation (Kumari et al., 2007; Serikawa et al., 2018). On the other hand, although 3' UTRs lie far from the translation initiation site, they can regulate translation efficiency by controlling polyadenylation and binding of *trans*-acting factors. For instance, the poly(A) tail of mRNA can loop with its 5' end, facilitating translation and ribosome recycling (Choe

et al., 2018; Wells et al., 1998). Moreover, proteins and miRNAs that are deposited on 3' UTRs can regulate various phases of translation including initiation, elongation, and termination (Mazumder et al., 2003).

Massively parallel translation assays

To decode the regulatory potential of UTRs, we developed a massively parallel reporter assays (MPRAs), in which UTRs of fixed length adjoin the coding sequence of a reporter. Similar approaches have been adopted to reveal the sequence-specific regulation and modular functions of numerous processes, including transcription, RNA splicing, translation, RNA localization, RNA decay and protein degradation in various species (Abell et al., 2022; Chiang et al., 2022; Griesemer et al., 2021; Jia et al., 2020; Litterman et al., 2019; Mikl et al., 2022; Niederer et al., 2022; Oikonomou et al., 2014; Rabani et al., 2017; Sample et al., 2019; Savinov et al., 2021; Slutskin et al., 2018; Zhao et al., 2014). In some studies of RNA translation, reporter protein output, for example, intensity of the green fluorescent protein (GFP), has been used as a readout of regulatory activity. However, protein output essentially is the product of the interaction between RNA half-lives and its translatability. Therefore, for a more unequivocal study design, poly(ribo)some profiling has been used to estimate RNA translatability through evaluating their ribosome association status. MPRA based on polyribosome profiling technique has been applied to examine the translational efficiency of UTRs of interest. For instance, it has been shown previously that variations in 5' UTRs have a greater impact on ribosome load than those in coding regions or 3' UTRs (Leppek et al., 2022). MPRAs on 5' UTRs have demonstrated that an upstream start codon (uATG) or open reading frame (uORF) can suppress primary ORF translation (Jia et al., 2020; Sample et al., 2019). Structural hindrances, such as RNA G-quadruplexes (RG4) and stem- loops, also can suppress translation (Fay et al., 2017; Jia et al., 2020; Leppek et al., 2022). Additionally, at the sequence level, G-rich and C-rich elements are repressors of translation, whereas A-rich elements enhance translation (Jia et al., 2020; Niederer et al., 2022). Most massively parallel polysome profiling studies exclusively examined 5' UTRs, likely due to the general assumption that 5' UTRs play the major role in ribosome recruitment. On the other hand, MPRAs assessing fluorescent protein expression showed that binding sites for miRNAs and RNA- binding proteins (RBPs), particularly AU-rich elements, played a critical role in regulating the protein output (Oikonomou et al., 2014; Stoecklin et al., 2001; Wissink et al., 2016). Pumilio and AU-rich element-binding proteins, such as Tristetraprolin (TTP) and the embryonic lethal abnormal vision (ELAV) family, have been shown to influence protein expression (Barreau et al., 2005; Mayya & Duchaine, 2019; Van Etten et al., 2012). Nevertheless, whether these RBP or miRNA binding sites dampened protein level via impeding ribosome association was not corroborated by those MPRA studies.

Precision diagnosis of UTR mutations

The genetic variants in UTRs may affect mRNA stability, translation, and localization, which in turn impact the flow of gene expression and hence result in phenotypical change and even pathogenesis. Prior studies have evidenced that even single nucleotide changes in a UTR can affect mRNA translation and/or transcript half-life in disease contexts. For example, single nucleotide substitutions at the 36th position of the 5' UTR or the 1723rd position and the 3' UTR of transforming growth factor β3 (TGFβ3; MIM# 190230) are associated with typical arrhythmogenic right ventricular cardiomyopathy (Beffagna et al., 2005). Another such example is a point mutation in the 3' UTR of glutamine-fructose-6-phosphate transaminase 1 (GFPT1), which results in congenital myasthenic syndrome. GFPT1 is a rate-limiting enzyme for hexosamine biosynthesis, and a mutation in its 3' UTR has been known to cause 90% reduction in protein production, potentially due to gain of a miRNA binding site (Dusl et al., 2015). All things considered, sequence integrity of UTRs has profound effects on gene regulation, given that variants in UTRs may induce various phenotypes and even lead to severe diseases.

To systematically evaluate the impact of disease-relevant UTR variants on translation, we employed massively parallel ribosome profiling on both 5' and 3' UTRs to examine the effects of mutations on ribosome association. Through this strategy, we identified 46 high-confidence variants that caused ribosome shifting, and, intriguingly, their functional consequences converged on changes of RNA secondary structure and RBD binding motifs. Moreover, we investigated the molecular mechanism of a severe 5' UTR suppressive mutation and developed ADAR-based editing to rescue the expression defect.

Material and methods

Metagene analysis

All UTR and uORF variants were assigned using dbSNP version 151 (Sherry et al., 1999), with disease variants and UTR sources as described above. Definitions of uORFs were downloaded from TIS-db (Wan & Qian, 2014), and the respective genome coordinates were lifted from the Genome Reference Consortium hg19 to hg38 before analysis.

Variant collection

Disease-related variants were collected from ClinVar (Landrum et al., 2016) and the HGMD database (Stenson et al., 2017). Variants located in the coding region or labeled as benign variants were excluded. UTRs were defined using NCBI RefSeq and ENCODE V27. Then, we intersected selected variant positions with the UTR regions by using BEDTools (Quinlan & Hall, 2010) to collect UTR variants.

Construction of a UTR library of DNA templates

A UTR library of DNA templates was assembled by overlapping extension polymerase chain reaction (PCR) using Herculase II Fusion Enzyme (Agilent Technologies). Initially, the oligonucleotide library sequences (CustomArray Inc., USA) were double-strandized and amplified by PCR. After PCR clean-up (Qiagen), the UTR library amplicons were appended sequentially with EGFP CDS and CMV promoter sequences (derived from EGFP-N1 vector) by means of overlapping PCR. The assembled full-length DNA templates (CMVP-5'UTR-EGFP or CMVP-EGFP-3'UTR) were subjected to PCR clean-up for subsequent transfection. Primers are listed in Supplemental Table S1.

Polysome fractionation and preparation for next generation sequencing

Polysome fractionation was conducted in accordance with a protocol published previously (Gandin et al., 2014) with slight modification. In brief, HEK293T cells cultured in 15-cm Petri dishes were grown to 60-70% confluency and then transfected with 4 μg of the UTR library of DNA templates (2 μg for each UTR library) using lipofectamine 3000 reagent (Invitrogen). Fourteen hours after transfection, the cells were incubated in culture medium containing 100 μg/ml cycloheximide (CHX) (Sigma) for 5 minutes. Subsequently, the cells were washed twice with and then detached in ice-cold 1x PBS containing 100 μg/ml CHX, followed by centrifugation at 200 x *g* for 5 minutes at 4 °C. The entire procedure was conducted at a low temperature with the addition of 100 μg/ml CHX to preserve the polysome status. After discarding the supernatant, the cell pellet was re-suspended with 425 μl of hypotonic buffer (5 mM Tris-HCl (pH 7.5), 2.5 mM $MgCl_2$, 1.5 mM KCl and 1x EDTA-free protease inhibitor cocktail (Roche)). Next, 5 μl of 10 mg/ml CHX, 1 μl of 1 M DTT, and 100 units of RNase inhibitor (Invitrogen) were added into the cell suspension, followed by vortexing for 5 seconds. Then, 25 μl of 10% Triton X-100 and 25 μl of 10% Sodium Deoxycholate were added into the suspension. After lysing the cells on ice for 5 minutes, the lysate was centrifuged at 16,000 x *g* for 7 minutes at 4 °C. The supernatant was collected and then layered

onto a 5-50% sucrose gradient made of 5% and 50% sucrose solutions containing 20 mM HEPES (pH 7.6), 0.1 M KCl, 5 mM $MgCl_2$, 10 μg/ml CHX, 0.1x EDTA-free protease inhibitor cocktail, and 10 units/ml RNase inhibitor, before undergoing ultracentrifugation with a SW41 Ti rotor (Beckman Coulter) at 36,000 rpm and 4 °C for 2 hours. Afterwards, the gradient samples were fractionated by fixed volume (500 μl/fraction), and all of the fractions were collected and subjected to RNA extraction (Direct-zol™ RNA MiniPrep Plus, Zymo Research). In addition to the indication of polysome profile derived from UV absorbance, the identity of each fraction was further verified through examining the distribution of 18S and 28S rRNAs among fractions via RNA electrophoresis.

Next, fractions corresponding to ribosome peaks were further categorized into three groups: monosome (Mono), low-density polysome (Light, 2-5 ribosomes), and high-density polysome (Heavy, ≥ 6 ribosomes). Then, the RNAs extracted from the fractions corresponding to each group were pooled together in an equal-fraction manner and subjected to reverse transcription using SuperScript IV First-Strand Synthesis System (Invitrogen). Thereafter, the UTR regions of mRNAs derived from our UTR library of DNA templates were amplified from the cDNA products using Phusion PCR Master Mix/GC Buffer (Thermo) and primers hosting Illumina sequencing adaptors and random tri-nucleotides for Illumina® NextSeq paired-end 150 sequencing.

RNA sequencing analysis

The Illumina NextEra adapter CTGTCTCTTATACACATCT and low quality bases (sliding mean quality < 20) were removed from the raw fastq using trimmomatic (Bolger et al., 2014). The qualified reads were then aligned to the reference library with hisat2 (Smith et al., 2017), and sorted by samtools (Danecek et al., 2021):

- hisat2 --mp 1,1 --rdg 1,1 -x indexFile -U fastq_R1.gz | samtools sort -o out_sorted.bam

After sorting, the “Exactly One Time Alignment” filter was applied to the results by samtools view command with the -q 60 parameter and the output index bam file was generated by samtools index command. A final raw count table was built using the *multicov* command in BEDTools (Quinlan & Hall, 2010).

Statistical method for detecting the effect of mutations on polysome profiles

To determine variants that changed polysome distribution, we used read counts in polysome fractions to build a negative binomial generalized linear model with a log link function for each pair of oligonucleotides as follows:

$$\ln(C_i) = \beta_0 + \beta_F F_i + \beta_E E_i + \beta_G G_i + \theta G_i F_i + \eta G_i E_i$$

where C_i are read count values of the ith replicate, F_i represents the three polysome fractions (Mono, Light and Heavy), E_i represents biological replicates, and G_i represents WT or mutant.

We tested the significance of $\theta\theta$ to determine the effect of the variant. To avoid potential batch effects, we further excluded candidates with significant $\eta\eta$ or β_E($p < 0.05$).

All statistical analyses were performed using R language (version 4.2.2). Linear models were constructed using the *glm.nb* function in R.

Construction of DNA templates for luciferase assay

Linear constructs (CMVP-5'UTR-FFLuc or CMVP-FFLuc-3'UTR) for luciferase assay were assembled by overlap extension PCR using Phusion DNA Polymerase (Thermo). Initially, DNA oligos (Integrated DNA Technologies Inc.) of selected UTR candidate sequences were double-strandized and amplified by PCR. Afterwards, the purified UTR amplicons were appended with firefly luciferase CDS (derived from pGL3-Basic vector) and the CMV promoter sequence (derived from EGFP-N1 vector) sequentially via overlapping PCR.

Luciferase assay, RNA extraction, and RT-qPCR

Fourteen to fifteen hours after co-transfection with 500 ng of linear constructs (CMVP-5'UTR-FFLuc or CMVP-FFLuc-3'UTR) and 100 ng of pRL-TK vector, HEK293T cells cultured in a 12-well plate were harvested and processed according to the instruction of Dual-Glo Luciferase Assay System (Promega) with minor modification. In brief, after removing media and washing once with PBS, the transfected cells were detached and then lysed in 180 µl of Dual-Glo reagent, before being subjected to moderate shaking on an orbital shaker at room temperature for 10 minutes. The lysate was then transferred to a new tube and underwent low-speed centrifugation to remove the cell debris. Then, 50 µl of clear supernatant was transferred to 96-well black plates to measure firefly luminescence with EnSpire Multimode Plate Reader (PerkinElmer) or SpectraMax Paradigm reader (Molecular Devices) and the software SoftMax Pro. Subsequently, 50 µl of Dual-Glo Stop & Glo Reagent was added into each well. After incubation at room temperature for 10 minutes, *Renilla* luminescence was measured as above.

Relative luminescence units (RLUs) were calculated by dividing the firefly luciferase readouts by those of the transfection control, i.e., *Renilla* luciferase. To evaluate translation efficiency, the RLUs were normalized to the transcript level measured by real-time PCR. Apart from those specifically annotated herein, all data is shown as the fold-change of the mutant relative to wildtype.

To calibrate the luciferase assay results against cognate mRNA expression level, total RNAs were extracted from the cells subjected to identical transfection conditions with TRIzol reagent (Invitrogen). Total RNA was extracted with Direct-zol RNA Miniprep kits (Zymo Research), before being subjected to reverse transcription. Then, cDNA was reverse-transcribed from 1 µg of RNA using 2.5 µM random hexamer with 1 µl (200U) SuperScript IV Reverse Transcriptase (Thermo Scientific). The reaction mixture was incubated at 24°C for 10 min to allow primer annealing, followed by incubation at 50°C for 10 min for reverse transcription. Subsequently, the reaction was terminated by incubating at 80°C for 10 min to inactivate the reverse transcriptase. Finally, the cDNA products were subjected to real-time PCR with Fast SYBR Green Master Mix (QuantStudio 12K Flex system, Applied Biosystems) to determine the expression level of firefly luciferase relative to that of *Renilla* luciferase.

Motif enrichment analysis

Motif analysis including RBP position weight matrix (PWM), K-mer and miRNA seed sequence. RBP PWM definitions were downloaded from the ATtRACT database (Giudice et al., 2016). For each UTR sequence, all possible motif PWMs were determined using the R package Biostrings (Pagès H, 2022) and the *matchPWM*() function with the parameter min.score = 95%. To summarize the contribution of each RBP, PWMs predicted for the same RBP gene were summed. K-mer composition was determined in Biostrings using the *oligonucleotideFrequency*() function, with K-mer lengths of 1 to 7. miRNA binding sites were predicted in TargetScan 7.0 (Agarwal et al., 2015), with high-confidence predictions (8mer, 7mer-m8, 7mer-1A) for each miRNA being summed to represent counts of the seed sequence.

To find motifs enriched in UTRs of interest, motif distribution in relation to UTRs was analyzed by Fisher's exact test using the *fisher.test*() function in R (R-Core-Team, 2021) according to the matrix:

$$\begin{bmatrix} count\ of\ motif[i] \in sig\ UTRs & count\ of\ motif[i] \in nonsig\ UTRs \\ all\ motif\ count \in sig\ UTRs & all\ motif\ count \in nonsig\ UTRs \end{bmatrix}$$

To examine motif changes in relation to polysome shifting, motifs disrupted by variants and the resulting shift in polysomes was analyzed by using the *fisher.test*() function in R according to the matrix:

$$\begin{bmatrix} sig\ UTRs\ with\ motif[i]\ change & nonsig\ UTRs\ with\ motif[i]\ change \\ sig\ UTRs\ without\ motif[i]\ change & nonsig\ UTRs\ without\ motif[i]\ change \end{bmatrix}$$

shRNA-mediated knockdown

For shRNA expression, pLKO.1 vectors containing an shRNA sequence (5'-AGTTGTGTAGCAGTTGAGTAA-3') targeting the human *SRSF2* 3' UTR or a non-targeting scrambled control (5'-CCTAAGGTTAAGTCGCCCTCG-3') were obtained from the RNAi Core, Academia Sinica, Taiwan, and lentiviral particles were produced using HEK293T cells by following standard procedures.

The HEK293T cells were initially seeded at a density of $5x10^5$ cells per well in 6-well plates. Four hours after seeding, shRNA-expressing lentivirus was introduced into the HEK293T cells at a multiplicity of infection (MOI) of 5. Then, the virus-containing medium was replaced after 24-hour incubation with 2 ml of DMEM supplemented with 10% FBS. Following an additional 24-hour incubation, the culture medium was supplemented with 2.5 μg/ml puromycin for selection purposes. The cells underwent a 4-day puromycin selection with one round of medium refreshment. The cells were then incubated in medium with 5 μg/ml puromycin for an additional day. After the selection process, the cells were transferred to a 10 cm dish for expansion and subsequent experiments.

Plasmid construction

For the firefly luciferase reporter plasmids, full-length UTRs were PCR-amplified from HEK293T complementary DNA (cDNA) with Phusion High-Fidelity DNA Polymerase (Thermo Scientific). Restriction enzyme cutting sites (*HindIII* and *NcoI* for 5 'UTRs, *XbaI* for 3' UTRs) were attached for digestion with Thermo Scientific FastDigest kit and for cloning with Rapid DNA Ligation kit (Thermo Scientific) into the pGL3-Promotor (Promega) backbone. Mutants were generated by mutagenesis PCR from wildtype constructs and the same cutting sites as for cloning were adopted.

For motif validation, the 5' UTR sequence of *MUTYH* (mutY DNA glycosylase:c.-43~-157) and the 3' UTR sequence of *MEFV* (Mediterranean fever:c.*17~131) were used, respectively, as the backbone of wildtype UTRs. Motifs were inserted into the *Xba1* site of the *MUTYH* 5' UTR and the *Nhe1* site of the *MEFV* 3' UTR.

For IRF6 ar151-expressing plasmids, the 151 bp antisense sequence with a mismatch in the center was PCR-amplified from wildtype luciferase plasmids. The PCR product was inserted into the pHE vector (BIOTOOLS) and transcription was driven by the human U6 promoter.

UTR secondary structure prediction

UTR secondary structures were predicted and plotted using the forna website (*http://rna.tbi.univie.ac.at/forna/*) (Kerpedjiev et al., 2015). Free energy was estimated using the viennaRNA package (Lorenz et al., 2011) and the *RNAfold* command with the default setting.

$$\begin{bmatrix} sig\ UTRs\ with\ motif[i]\ change & nonsig\ UTRs\ with\ motif[i]\ change \\ sig\ UTRs\ without\ motif[i]\ change & nonsig\ UTRs\ without\ motif[i]\ change \end{bmatrix}$$

Elastic net model of polysome-shifting UTR variants

To develop a prediction model for determining whether a genetic variant affects mRNA translation, we employed a logistic regression model with elastic net regularization for feature selection. The curated features included free energy, structure rate, GC content, k-mer, G-quadruplex (RG4), conservation level, miRNA binding motif, PUM motif, RBP motif, and ARE.

Before feature selection, the features were narrowed down using multiple logistic regression models to regress the significance of variants (Yes/No) on each feature, with a *p-value* < 0.01 serving as the criterion for identifying potential candidates. Subsequently, the dataset was divided into two groups, comprising training and testing sets in a 2:1 ratio, respectively. The model is represented as follows:

$$\ln\left(\frac{\mathrm{P}(Y_i=1)}{1-\mathrm{P}(Y_i=1)}\right)=\beta_0+\sum_{j=1}^{p}\beta_j X_{ji}$$

where Y_{ii}= 1 is the ith mt sequence that significantly affects mRNA translation, and XX_{jjii} is the feature *jj* in the *m* th sequence. The predictive features were standardized using Z-score transformation to achieve unit standard deviation, thereby ensuring comparability of coefficients.

Elastic net regularization minimizes $L_{\log}+\lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{p}\beta_j^2+\alpha\sum_{j=1}^{p}|\beta_j|\right)$, where the negative log likelihood is:

$$\mathrm{L}_{\log}=-\sum_{i=1}^{n}\left[-\ln\left(1+e^{\left(\beta_0+\sum_{j=1}^{p}\beta_j X_{ji}\right)}\right)+y_i(\beta_0+\sum_{j=1}^{p}\beta_j X_{ji})\right]$$

We applied 10-fold cross-validation on the training set to optimize the tuning parameters (λ and α) for the elastic net model. Following this, predictions were made on the testing set to evaluate the model's performance, and the area under the receiver operating characteristic (AUROC) was calculated. Feature selection and AUROC calculations were conducted using the R packages 'caret' (v.6.0.94) and 'precrec' (v.0.14.4) (Kuhn, 2008; Saito & Rehmsmeier, 2017), respectively, with all analyses performed on R version 4.3.3. The 95% confidence intervals for the AUROC values were estimated through bootstrap resampling with 1,000 iterations.

Validation of the elastic net model

To further validate the predictive performance of our elastic net model, we conducted external validation using an independent dataset published by Sample et al. (2019). This dataset comprises curated human 5' UTR sequences that were systematically evaluated using a massively parallel translation assay to assess their effects on translation efficiency (Sample et al., 2019). The dataset was obtained from the authors' publicly available repository (*https://github.com/pjsample/human_5utr_modeling*).

Prior to analysis, rigorous quality control measures were implemented to ensure data reliability. Sequences were included only if they met the following criteria: an average read count of at least 50 reads per polysome fraction and a minimum of 10 reads in each individual fraction. Since our predictive model outputs binary classifications (significant effect versus no significant effect), we established a threshold to categorize the continuous MRL change values from the Sample et al. dataset. Variants with an absolute $\log_2$ mean ribosome loading (MRL) change of ≥1.0 (corresponding to the "log_obs_diff" column in their dataset) were classified as having significant translational impact, while those below this threshold were considered non-significant. Model performance on the external dataset was evaluated by calculating the AUROC with 95% confidence interval using the same methodology described previously.

It should be noted that external validation for the 3' UTR predictive model could not be performed due to the absence of suitable independent datasets with comparable experimental design and quality standards.

Western blotting

To validate translation activity on the mutant-generated upstream AUG, HEK293T cells were cultured in 6-well plates. After 24 hours, the cloned reporter plasmids and the control pEGFP-N1 were co-transfected by Lipofectamine 3000 (Invitrogen) at a ratio of 3:1 (2.5 μg in total). At 24 h post-transfection, cells were lysed in RIPA buffer (Sigma-Aldrich) supplemented with cOmplete Protease Inhibitor Cocktail (Roche). The solution was transferred onto ice for incubation for 5 minutes and then centrifuged at full speed for 10 minutes at 4 °C. Finally, the supernatant was transferred to a new tube. After determining protein concentration by Bradford assay (Bio-Rad), 30-80 μg of protein sample was loaded into each well. Protein samples separated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis were transferred to PVDF membrane for antibody incubation and ECL detection. Immunoblotting of IRF6: primary antibody against IRF6 (Cell Signaling Technology, Cat# 6948, dilution: 1:1000) and secondary antibody (Jackson ImmunoResearch, Cat#: 111-035-003, 1:5000); Immunoblotting of tubulin: primary antibody against tubulin (Sigma, Cat#: T6793, 1:2000) and secondary antibody (Cell Signaling Technology, Cat#: 7076, 1:5000); Immunoblotting of firefly luciferase: primary antibody against luciferase (Abcam, Cat#: ab185924, 1:1000) and secondary antibody (Abcam, Cat#: ab6721, 1:5000); Immunoblotting of GFP: primary antibody against GFP (ProteinTech, Cat#: HRP-66002, 1:10000).

ADAR editing

HEK293T cells were cultured in 24-well plates. After 24 hours, 200 ng of the reporter plasmids, 20 ng of the pRL-TK plasmids, 150 ng of the ar151 plasmids, and 150 ng of the ADAR2 overexpression plasmids were co-transfected with Lipofectamine 3000 (Invitrogen). Luciferase assays and evaluations of mRNA levels were performed as described above after 40 hours of transfection.

To evaluate the editing efficiency of IRF6 ar151, the cDNA synthesized for qPCR analysis was PCR-amplified with primers targeting 151 bp of the 5' UTR. PCR products were purified with QIAquick PCR Purification Kit (QIAGEN) according to the manufacturer's protocol. Elution products were subjected to Sanger sequencing to analyze editing efficiency.

Results

Meta-analysis of disease-relevant UTR variants

It has been estimated that ~3.7% of disease-relevant genetic variants arise in UTRs (MacArthur et al., 2017; Steri et al., 2018). To investigate if there is a positional effect of disease-relevant UTR variants, we compared the overall distribution of disease-associated single nucleotide polymorphisms (SNPs) in ClinVar (Landrum et al., 2016) to over a million human UTR SNPs in the dbSNP archive, which revealed enrichment for disease-relevant variants near mRNA coding regions in both 5' and 3' UTRs (**Fig. 1A**), indicating that the near-coding region of UTRs is a critical regulatory region for gene expression. Although we did not detect enrichment for disease-relevant 5' UTR variants near the transcription start site, they did exist at uORF start sites (**Fig. 1B**). This specific localization of disease-relevant UTR variants evidences a trend in the positioning of functional elements within UTRs.

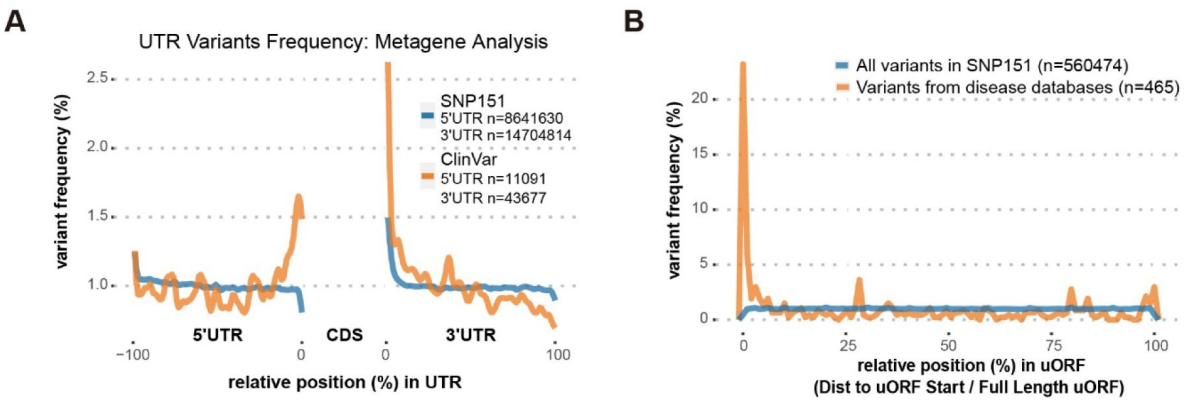


Figure 1.

Enrichment of UTR variants relative to open reading frames.

(A) Metagene analysis of native human UTRs (blue line) or disease-relevant variants reported in ClinVar (orange line). Note the enrichment of UTR variants near the coding region. **(B)** Metagene analysis of all native human variants (blue line) or disease-relevant variants reported in ClinVar and HGMD (orange line) in relation to uORFs. Note the distinct enrichment of disease-relevant variants at the uORF start site.

Massively parallel poly(ribo)some assays on human disease-relevant UTR mutations

Although many disease-associated UTR variants are located in functional elements, their pathogenic mechanisms remain unknown. To examine if these variants directly affect protein output, we first sought to determine the impact of these variants on translation. To systematically examine the altered translation efficiency by UTR variants, we implemented massively parallel polysome assays to examine the effects of variants on polysome association (**Fig. 2A**). First, we extracted 6,555 disease-associated UTR variants, including SNPs and indels shorter than 15 nucleotides (nt), from the Human Gene Mutation Database (HGMD) and ClinVar, followed by multiplex oligonucleotide synthesis for selected UTRs. We prioritized UTR variants designated as pathogenic or likely pathogenic in the database. Consequently, all annotated pathogenic and likely pathogenic UTR variants reported by the time of data collection were included in the library. Additionally, due to the constraint of synthesis length (~150 nt), we synthesized the segment of selected UTRs (~115 nt) in which variants were located at the center to maximize the possibility of including critical regulatory elements. Moreover, UTR fragments of both alternative and reference alleles were synthesized in parallel for case-control comparison. In cases where multiple alternative alleles existed at the same position, they were paired with the same reference allele to prevent duplication in the library. The entire library of UTR oligonucleotides (UTR library) was subsequently ligated upstream or downstream of an enhanced GFP (EGFP) coding region, along with a CMV promoter and a common UTR sequence on the opposite end. Cells transfected with the UTR library were treated with cycloheximide 14 hours post transfection and then subjected to polysome fractionation (see Methods). By matching the UV absorbance profile of fractions to rRNA abundance revealed by gel electrophoresis of RNA samples (total RNA extracted from polysome fractions), we were able to estimate the number of bound ribosomes in the collected fractions. RNA samples from fractions containing 1 ribosome, 2 – 5 ribosomes, and 6 or more ribosomes were pooled, respectively, and are referred to hereafter as the Mono, Light and Heavy groups.

UTR library constituents were then reverse-transcribed and PCR-amplified from the three groups of RNA samples, before undergoing next-generation sequencing. To evaluate the effects of disease-associated variants on the distribution of corresponding transcripts across the Mono, Light and Heavy groups, we fit a negative binomial generalized linear model to read count for each UTR pair (wildtype/WT and mutant/mut) in three fraction groups while accounting for batch effects of the three experimental replicates (see Methods). The model primarily focused on discerning the distribution pattern of reads in three polysome fractions between UTR pairs, without considering differences in the expression levels of the UTR species. As a result, the analysis was unaffected by any perturbation of UTR sequence at the transcriptional level and did not account for changes in stoichiometry between UTRs and regulatory factors due to variations in expression levels. Consequently, we identified 483 high-confidence UTR pairs, comprising 396 pairs that did not alter translation (172 and 224 for 5' UTRs and 3' UTRs, respectively; collectively denoted by high-confidence nonsignificant or 'HC-nonsig' UTR variants), and 19 5' UTR and 27 3' UTR variants that significantly affected polysome distribution (**Fig. 2B**, Supplemental Figs. S1, S2 and Table S2; collectively denoted by high-confidence significant or 'HC-sig' UTR variants). Pearson correlation analysis revealed R coefficients ranging from 0.59 to 0.71 for the mut-to-WT transcript ratios across three independent experiments (Supplemental Fig. 3).

Although a very small proportion of UTR variants are categorized as pathogenic, we observed that HC-sig variants are enriched with known pathogenic variants relative to the HC-nonsig variants (**Fig. 2C**), suggesting a contribution of translation-altering UTR variants to pathogenicity. Notably, we found that these translation-altering UTR variants are particularly prevalent in neurological and developmental disorders (Supplemental Table S3).

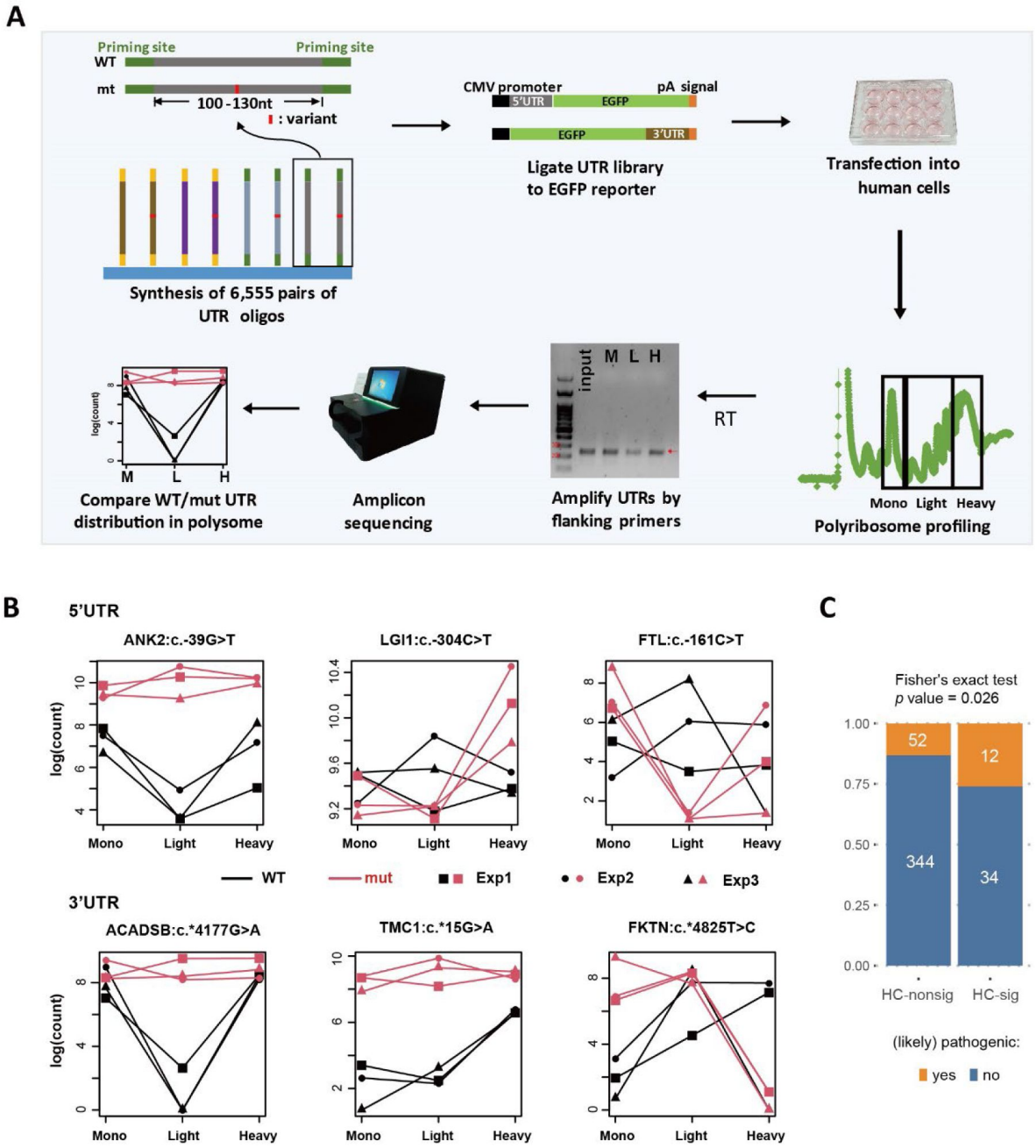


Figure 2.

Massively parallel polysome assays.

(A) UTR fragments were synthesized in bulk and ligated with EGFP reporters for expression in human cells. The reporter-expressing cell lysate was fractionated into monosome, light (2-5 ribosomes) and heavy (6+ ribosomes) polysome- containing fractions, and the UTR library was then amplified for amplicon sequencing. Mutations that altered relative distribution among fractions were defined as significant polysome-shifting mutations. **(B)** Examples of UTR pairs that significantly differed in polysome distribution. **(C)** Enrichment of pathogenic variants among HC-sig polysome-shifting UTR variants.

Even with the high confidence dataset, we did not intend to immediately translate the altered polysome profile into an increase or decrease in translation efficiency, as the direction of the shift was not readily evident. Additionally, sedimentation in the sucrose gradient may have been partially affected by heavy particles other than ribosomes. Subsequently, individual reporters were used to examine the variants' effect on translation.

Validating the translation defects of mutant UTRs by luciferase reporter assay

The variants significantly altering the polysome profile were then individually validated through high-sensitivity luciferase reporter assays (**Fig. 3A**). To this end, we resynthesized both the variant and corresponding wildtype alleles in the same library format - 115-nt native UTR segments centered on the variant and flanked by 20-nt priming sites. These UTRs were then cloned upstream (5') or downstream (3') of the firefly luciferase coding sequence, depending on their genomic location. As the initial library design, the test UTR segment differs only by one nucleotide, while a shared short UTR fragment is present on the opposite end of the coding sequence to ensure consistency across constructs (**Fig. 2A**). To compare translation activity between wildtype and mutant UTRs, two reporter constructs of the same pair were transfected into independent wells of HEK293T cells, whereby their bioluminescence intensity was determined separately. Notably, to calibrate transfection efficiency across wells, the UTR reporter constructs were co- transfected with *Renilla* luciferase vector, and the firefly bioluminescence signal intensity was normalized to that of *Renilla*. Moreover, to control for potential variations in transcriptional activity, the relative firefly luminescence (mut/WT) of each UTR pair was further normalized to their cognate relative RNA level (mut/WT) (see Method). Our results indicate that 11 of the 12 selected HC-sig variants displayed prominent effects on translation activity (**Fig. 3A**). Additionally, most of the tested UTR variants only altered translation efficiency but not RNA level, except for RFAXP:c.-101T>G (Supplemental Fig. S3). Furthermore, we observed that 5' UTR variants had a greater impact on translation activity relative to 3' UTR variants (**Fig. 3C**).

After validating individually the variants identified in our polysome profiling assay, we proceeded to investigate their effects on translation within an endogenous genomic context, targeting a total of eight variants (five in 5' UTRs and three in 3' UTRs). We derived the respective full-length UTR sequences from HEK293T cells and introduced the variants into cognate UTR amplicons via site-directed mutagenesis. Next, the UTR fragments were cloned into firefly luciferase reporter plasmids, which were then used for luciferase assay. 15 hours post transfection, the translation efficiency of the full-length UTR constructs was determined and normalized as described above. Despite being embedded in a longer genomic context, 7 of the 8 selected variants still significantly altered the translation activity of the full-length UTR constructs (**Fig. 3B**). Moreover, we observed only minor variation in the mRNA levels of the luciferase reporter across the full-length UTR constructs, indicating that the observed phenotypes could not be attributed to discrepancies in transcriptional activity or RNA stability (Supplemental Fig. S4). Furthermore, UTR variant-elicited regulation was highly consistent between full-length and short-form UTRs (**Fig. 3C**), though the fold-change (mut/WT) in translation activity observed for full-length UTRs was much lower than for short-form UTRs, potentially due to the existence of additional regulatory elements in genomic milieus, which hindered or mitigated the impact of variants. Thus, our luciferase reporter assays verify the robustness of our massively parallel polysome assay results. The disparity between full- length and short-form UTRs also highlights the strength and weakness of our strategy.

Polysome-shifting UTR variants promote motif gain/loss

We hypothesized that polysome-shifting UTR variants may result in the removal or gain of functional motifs that recruit *trans* factors responsible for regulating translation. Accordingly, we postulated that polysome-shifting variants would display distinct sequence properties. Therefore, first we investigated the nucleotide composition of polysome-shifting variants. We found that

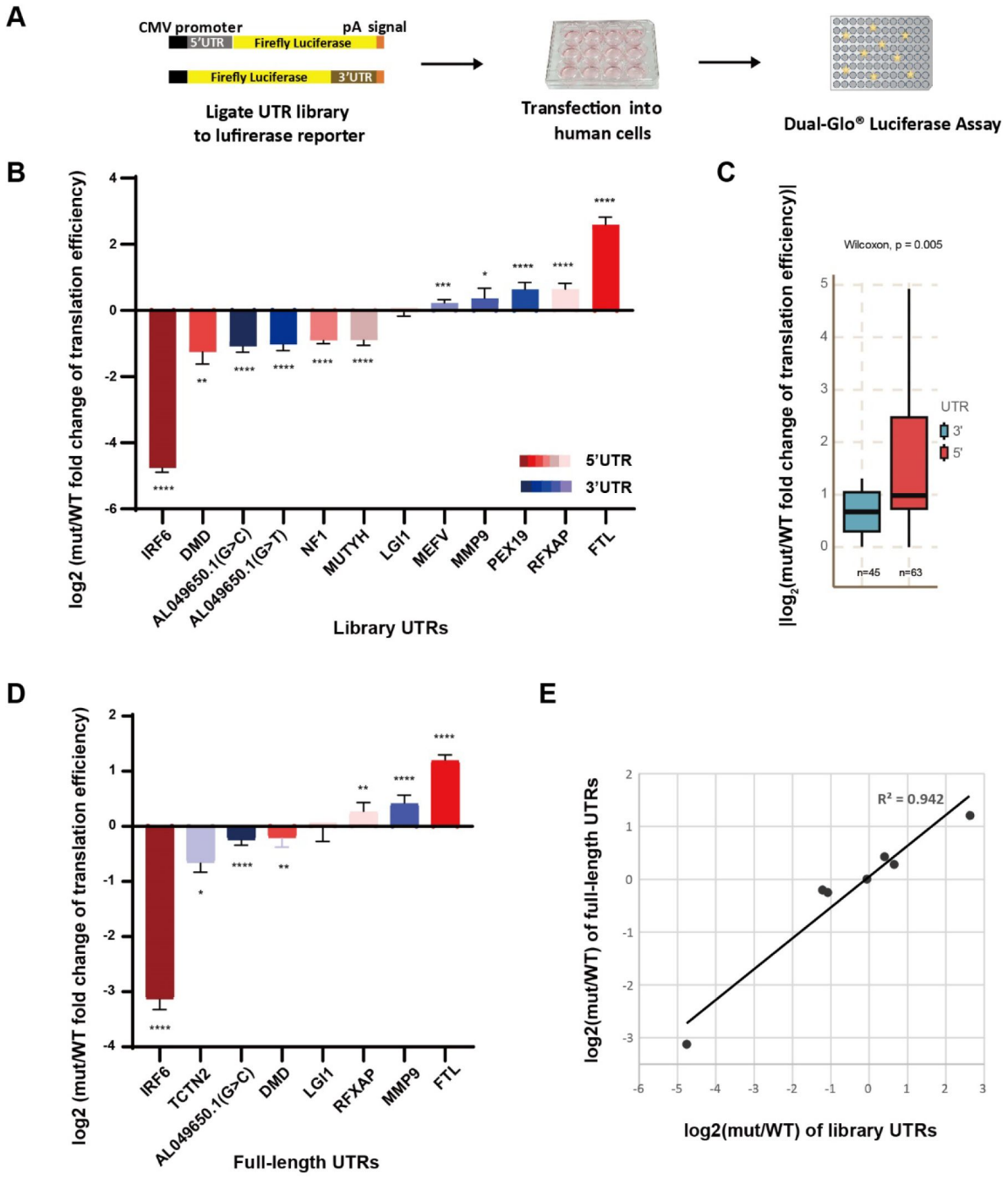


Figure 3.

Validation of mutant effect on translation by luciferase assays.

(A) Selected UTRs were ligated with firefly luciferase reporters and co-transfected with *Renilla* luciferase into human cells. The firefly luminescence was first normalized to the *Renilla* luminescence, and further normalized to the respective RNA expression levels to determine translation efficiency. The log2 value of mut over WT translation efficiency was plotted ascendingly. **(B)** Translation efficiency of eight full- length UTR pairs was examined in the same manner as in (A). UTR pairs are color-coded according to their gene origin (A&B). The individual assays were repeated three times and the differences were accessed by a two-tailed Student t-test: *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$ (A&B). **(C)** Absolute log2 translation efficiency changes (mut vs. WT) were compared between 5' and 3' UTR variants, as shown in (B). Statistical significance was assessed using a two- sided Wilcoxon rank-sum test. **(D)** Comparison of the effect of mutations on the translation efficiency of library (short) and full-length forms of UTRs.

variations in C nucleotides (including gain and loss) in 5' UTRs were relatively inert in terms of translational alteration, whereas that was not the case for C in 3' UTRs (Supplemental Fig. S5A and Table S4). Subsequently, we investigated the RBP-binding motifs of polysome-shifting UTR variants. Our results show that GAAGAA and T-rich motifs are enriched in HC-sig 5' UTRs, whereas G-rich and YTTC motifs are enriched in HC-sig 3' UTRs (Supplemental Fig. S5B and Table S4, Y represents T or C). Furthermore, recognizing the intricate relationships between binding motifs and their corresponding RBPs, we conducted an additional analysis focusing on the RBPs themselves. This analysis suggested that TRA2B and members of the HNRNP family can bind HC- sig UTRs to regulate their translation (Supplemental Fig. S5C). Extending this analysis, we calculated if changes in RBP-binding motifs are associated with changes in polysome profiles, which revealed that changes of SRSF1/2-, PABPN1- (AGAAG motifs) and CELF2- (T-rich motifs) binding sites in 5' UTRs frequently caused polysome shifting, so did mutations of binding sites for KHSRP (AGGG motifs) in 3' UTRs (**Fig. 4A**).

To further validate the RBP-binding motifs identified by our enrichment analysis, we examined the effect of selected motifs on reporter expression by inserting them into a segment of the *MUTYH* 5' UTR or *MEFV* 3' UTR that contains suitable cloning sites (see Methods) and then perturbing its effect by mutating it so that it retained the same length but rendered it non-RBP-binding according to the ATtRACT RBP motif database (Giudice et al., 2016). For 5' UTR, GTTTTG and GAAGAAG motifs were selected for the test. Both selected motifs significantly increased the translation efficiency of luciferase reporters without affecting RNA expression level (**Fig. 4B**). For 3' UTR CTTTTC and AGGGGGA motifs were selected for the test, yet neither of them significantly altered reporter expression. However, mutation of the AGGGGGA motif into AGCTGTA significantly reduced translation efficiency, indicative of a context-dependent effect of RBP-binding motifs (**Fig. 4C**). This observation is consistent with additional findings where variants that create or disrupt specific RBP binding sites—such as SRSF1/2 (e.g., in *DMD* and *NF1*; **Fig. 3** and Supplemental Fig. S4) and KHSRP (e.g., in *AL049650.1*; **Fig. 3** and Supplemental Figs. S4 & S5)— led to significant changes in translation efficiency within their native UTR contexts.

Additionally, to assess the impact of RBPs on translation regulation through the short repeat sequence, we conducted a knockdown of SRSF2 using shRNA. This RBP was chosen due to its cytoplasmic localization (Thul et al., 2017) and abundant expression in HEK293T cells compared to other candidate RBPs. The results revealed that depleting SRSF2 eliminated the translation enhancement mediated by GAAGAAG, while having no effect on a non-SRSF2-binding motif (**Fig. 4D**). While SRSF2 has not been direcr1tly implicated in translation regulation, some members of the SRSF family have been reported to function as translation regulators (Sliskovic et al., 2022). Notably, a mutation in the *NUMA1* 5' UTR that disrupted SRSF9 binding was previously demonstrated to decrease translation efficiency (Lim et al., 2021), providing further support for the involvement of SRSF family proteins in translation regulation. Together, these results support the existence of mutational hotspots on short repeated sequences for polysome-shifting UTR variants and a collaborative regulatory effect of binding motifs.

Polysome-shifting mutations alter RNA secondary structure

Given that RNA secondary structure is believed to control translatability in addition to primary sequences (Georgakopoulos-Soares et al., 2022; Leppek et al., 2018), we determined the changes in folding energy elicited by sequence variations. To do so, we compared the folding energy change for variants in the HC-sig and HC-nonsig groups, which revealed that the 5' UTR HC-sig variants significantly weakened the folding structures and tended to prompt greater structural change than HC-nonsig variants, which was not the case for 3' UTR variants (**Fig. 5A**). Furthermore, the HC-sig WT exhibited robust folding energy, potentially underlying their susceptibility of folding perturbations. In the same vein, RNA secondary structure predictions showed notable structural changes caused by 5' UTR single-nucleotide variations. As an example

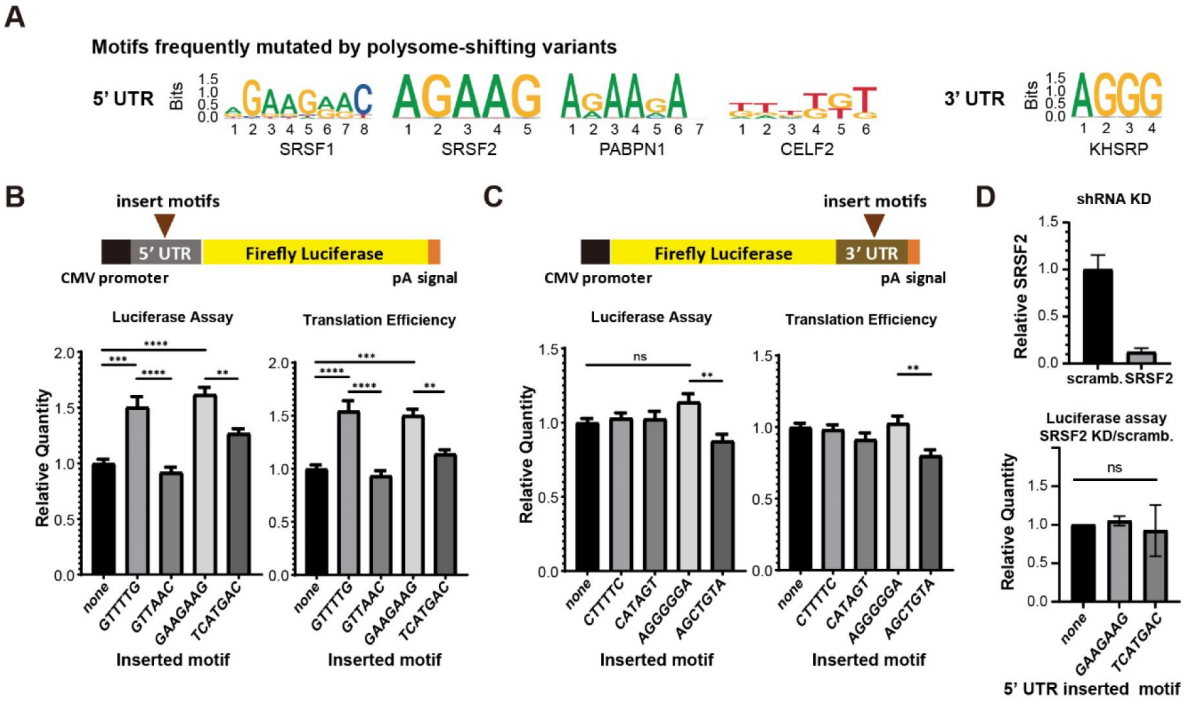


Figure 4.

Identification of polysome-shifting RBP-binding motifs.

(A) Motifs significantly more frequently mutated by HC-sig variants. **(B)** Two 5′ UTR-enriched motifs from (A), i.e., GTTTTG and GAAGAAG, were inserted into a 5′ UTR of firefly luciferase to determine their luciferase activity and normalized translation efficiency. Mutant forms of these two motifs, GTTAAC and TCATGAC, were generated for comparison. **(C)** Similar to the 5′ UTR design in (B), the 3′ UTR-enriched motifs CTTTTC and AGGGGGA were examined against their mutant forms, i.e., CATAGT and AGCTGTA. The individual assays were repeated four times. Statistical significance was determined by a two- tailed Student t-test: **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$. **(D)** Upper: Relative SRSF2 transcript levels, as determined by RT-qPCR. Lower: Relative luciferase activity with various 5′ UTR motifs, normalized to scrambled shRNA infection.

of the latter, we present the case of a 5' UTR single nucleotide variation in C-C chemokine receptor type 5 (CCR5) in **Figure 5B**. These results suggest that structural alterations of 5' UTRs by sequence variation could be a potential causal factor of the translation defects.

Since it has been shown previously that the usage of an uATG is affected by the surrounding UTR structure and can regulate translation of primary ORFs (Guenther et al., 2018; Xiang et al., 2023), we examined if the significant structural change elicited by HC-sig variants could be associated with the presence of uATG. Indeed, by contrasting UTR variants that did not alter uATG counts, we found that uATG-hosting UTRs were strongly associated with structural changes (**Fig. 5C**). Therefore, while changes in uATG may not be common explanatory factors for polysome-shifting mutations, our results suggest that structure-modifying UTR variants may control primary ORF translation partly by interfering with translation initiation from a uORF.

Modeling the polysome-shifting UTR variants

Recognizing that polysome-shifting variants encompass diverse classes, such as those disrupting RBP binding or secondary structure, we employed statistical learning to discern the contributions of sequence and structural features to polysome shifting. Over 3,000 features, including evolutionary conservation, folding energy, secondary structures, RBP motifs, and k-mers, were collected. Features deemed insignificant or with an area under the receiver operating characteristic (AUROC) smaller than 0.5 by univariate logistic regression were filtered out (Methods). The remaining features were further refined using elastic net regression. Twelve features were selected to elucidate polysome-shifting variants in the 5' UTR, while eighteen were chosen for the 3' UTR (**Fig. 6A** and Supplemental Table S5). Notably, GAA-repeat motifs of SRSF family binding sites again emerged as the most influential sequence feature for the 5' UTR. Additionally, changes in 'CAT' and 'AA' sequences, along with the folding energy of the 5' UTR, proved most effective in inducing polysome shifting. In contrast, for the 3' UTR, the structural contribution was less pronounced. Mutations altering GA- and T-rich sequences were identified as the strongest features promoting polysome shifting, consistent with prior analyses of RBP binding motifs (**Fig. 4**). Model accuracy was validated against a test dataset not included in the training, with both the 5' UTR and 3' UTR models achieving an area under the receiver operating characteristic curve (AUROC) exceeding 0.8, underscoring the model's robustness (**Fig. 6B**). To further assess generalizability, we applied the model to an external dataset that independently measured the effects of 5' UTR variants on polysome association (Sample et al., 2019). The model achieved an AUROC of 0.75 on this external benchmark (**Fig. 6C**), affirming its predictive power. Overall, our model supports the hypothesis that GA- and T-rich short repeat motifs serve as mutation hotspots for polysome-shifting mutations, while secondary structure exerts a greater influence on 5' UTR translation regulation compared to the 3' UTR.

The translation defect of a uORF-creating variant can be rescued by RNA editing

While uATG in the 5' UTR was not identified by the model as a widespread feature explaining polysome shifting, we hypothesize that it may still be a crucial factor. Its significance, however, might not have been universally captured across our library. Genome-wide analysis clearly indicated an enrichment of uORF disruption in the pathogenic UTR mutations (**Fig. 1B**). Therefore, we aimed to dissect the regulatory mechanisms of uORFs underlying the effects of polysome- shifting variants and develop rescue strategies against pathogenic uORF mutations.

First, we examined whether the presence of uATG in general leads to lower translation in reporter assays. We observed that the presence of uATG in 5' UTRs of firefly luciferase was correlated with lower protein outputs (**Fig. 7A**, Supplemental Table S6). Furthermore, two of the validated UTRs —Ferritin light chain (FTL) and Interferon Regulatory Factor 6 (IRF6)—demonstrated extreme translation perturbations (**Fig. 3**). In a study by Luscieti et al., an FTL variant (c.-161C>T) was

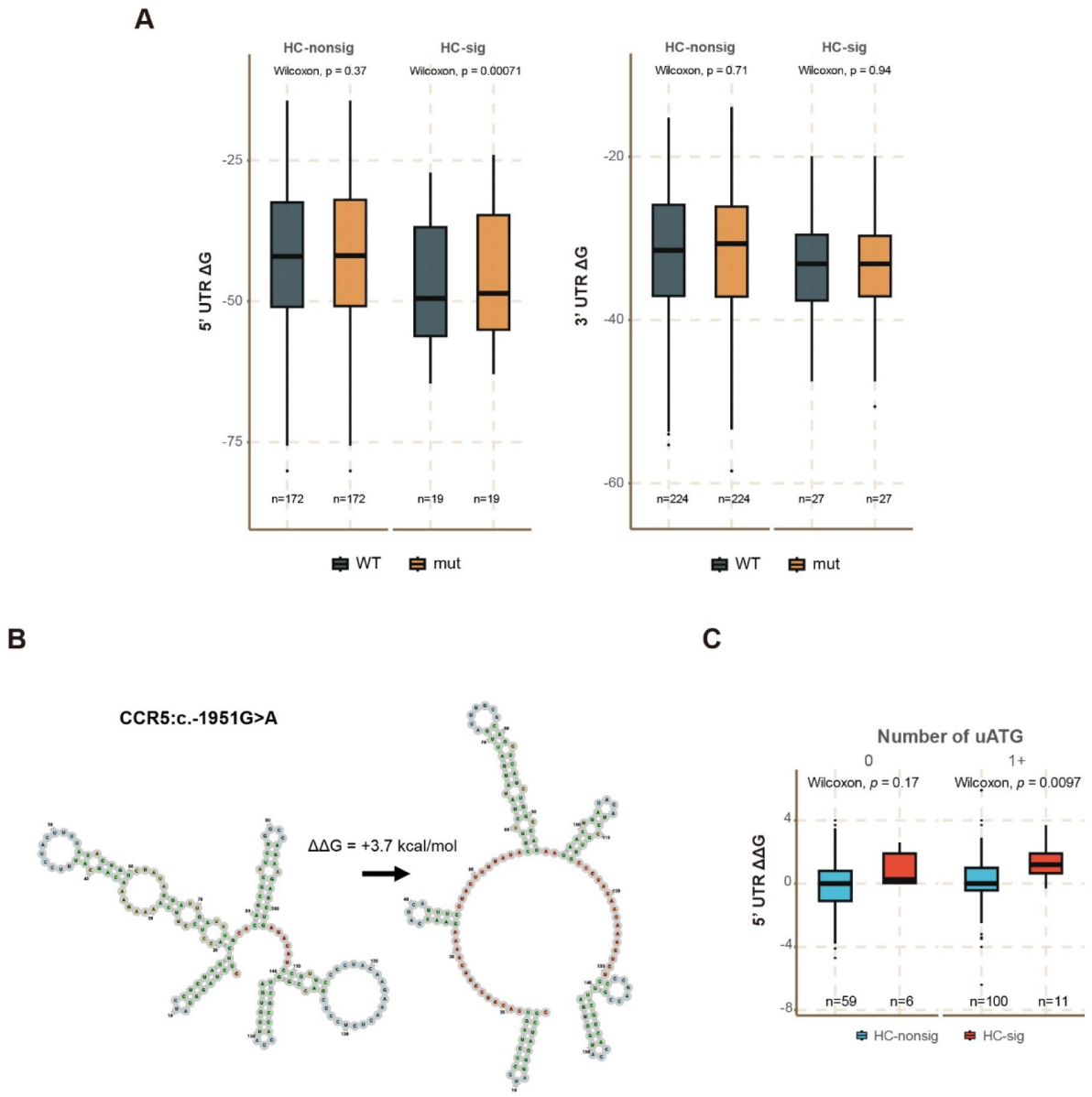


Figure 5.

5' UTR Polysome-shifting mutations elicit structural changes.

(A) Significantly weaker folding energy of 5' UTR polysome-shifting variants. The left panel is 5' UTR ΔG and the right panel is 3' UTR ΔG. **(B)** Structural change induced by a CCR5 5' UTR single nucleotide change, as determined by RNAfold. **(C)** Significant differences in ΔΔG values for 5' UTRs that contain upstream ATG (uATG).

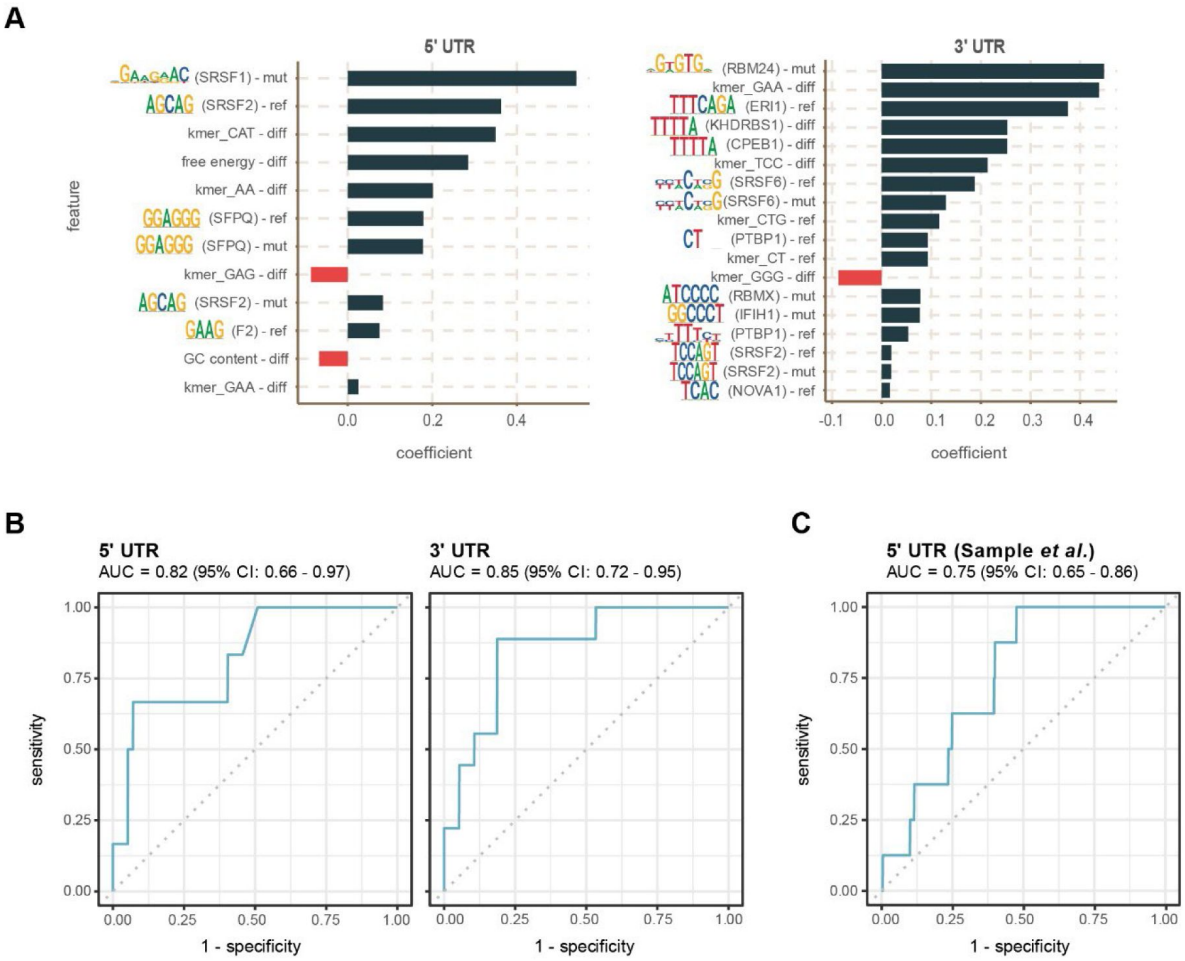


Figure 6.

Elastic net model distinguishes polysome-shifting UTR variants.

(A) Selected features from elastic net regression explaining the likelihood of a polysome-shifting UTR variant. A positive coefficient indicates a feature that promote polysome shifting, while a negative coefficient suppresses it. “mut” represents a feature of the mutant UTR, “ref” represents a feature of WT UTR, and “diff” indicates the difference (delta) associated with the variant. **(B)** Receiver operating characteristic (ROC) curve for the prediction of polysome-shifting variants on the test set of UTR variants. **(C)** ROC curve for the prediction of polysome-shifting variants on an external dataset of UTR variants (Sample et al., 2019).

characterized as preventing 5' UTR stem-loop formation (Luscieti et al., 2013). As a consequence, the recruitment of suppressive iron regulatory proteins mediated by the 5' UTR stem-loop is disrupted, leading to an increase in translation activity. An IRF6 variant (c. -4609 G>A) was found previously to be associated with the developmental disorder Van der Woude syndrome (de Lima et al., 2009). The G>A variant was predicted to generate an out-of-frame uORF, but no experimental evidence has been provided yet. As shown in **Figure 3**, the corresponding full-length mutant UTR displayed a dramatic 9-fold reduction in translation efficiency, strongly indicating that presence of the uORF represses primary ORF translation. To verify the role of uATG, we performed western blotting to measure protein levels directly and determine protein size. To do so, we generated a third reporter construct that harbored the uORF in-frame to the coding region by deleting one nucleotide (**Fig. 7B**). When probed with a luciferase-targeting antibody, the signal intensity for the out-of-frame mutant after normalization to the wildtype was only 0.089 (**Fig. 7C**). This suggests that only a minute fraction of ribosomes may reinitiate at the canonical site or undergo leaky scanning at the uATG. Notably, the in-frame mutant, which represents a combination of the extended and the original forms, displayed a 0.24-fold signal intensity, indicating that uATG, despite lower initiation efficiency, markedly suppresses initiation at the canonical site. Compared to wildtype, we determined 0.6-fold translation efficiency, as measured by luciferase reporter assays, upon overexpression of the in-frame mutant (**Fig. 7D**) and no significant change in mRNA level was observed (Supplemental Fig. S4). The discrepancy between protein abundance and luciferase intensity indicates that the extended form of the luciferase reporter may contribute to higher luminescence intensity.

To further validate the endogenous effect of the novel upstream ATG (uATG), we generated CRISPR knock-in clones carrying the IRF6:c.-4609G>A variant and examined its impact on gene expression. The introduction of the uATG reduced RNA levels to 88% and 37% of the wild-type in two independent clones (**Fig. 7E**), and protein levels to 44% and 23%, respectively (**Fig. 7F**), resulting in an overall reduction of translation efficiency to 50–62%. Although the uORF-mediated impairment of translation efficiency in IRF6 was clearly validated, we were unable to detect the corresponding micropeptide in the out-of-frame constructs (data not shown). It remains unclear if micropeptides, if translated, play any role in pathogenesis. However, patients who have gained a uORF in the IRF6 5' UTR phenocopy Van der Woude patients hosting loss-of-function IRF6 mutations (de Lima et al., 2009).

Lastly, we aimed to develop rescue strategies that may be useful for future therapeutic purposes. We selected site-directed RNA editing using adenosine deaminase acting on RNA (ADAR) enzyme for this purpose to avoid genome editing and potential off-target effects. Upon binding of an antisense RNA to a target transcript, the ADAR enzyme catalyzes adenosine-to-inosine editing. Inosine is subsequently interpreted as guanosine during translation (Merkle et al., 2019; Qu et al., 2019). To achieve high editing efficiency, we overexpressed the antisense RNA coding construct and ADAR2 protein (**Fig. 7G**). The transcribed antisense RNA encompasses a T>C mismatch that signals the complementary nucleotide for editing. First, we assessed editing efficiency by means of Sanger sequencing (**Fig. 7H**). Only co-expression of antisense RNA and ADAR2 resulted in detectable editing, with an efficiency of ~21%. Luciferase signal intensities were also validated (**Fig. 7I**). Consistently, expression of the antisense RNA alone had a very limited effect, but co- introduction of ADAR2 significantly increased the protein expression to ~50% the level of wildtype. Our promising results advocate the potential of clinical use of guided ADAR-RNA editing to reverse uORF-mediated translation suppression and thus alleviate the impact of the disease.

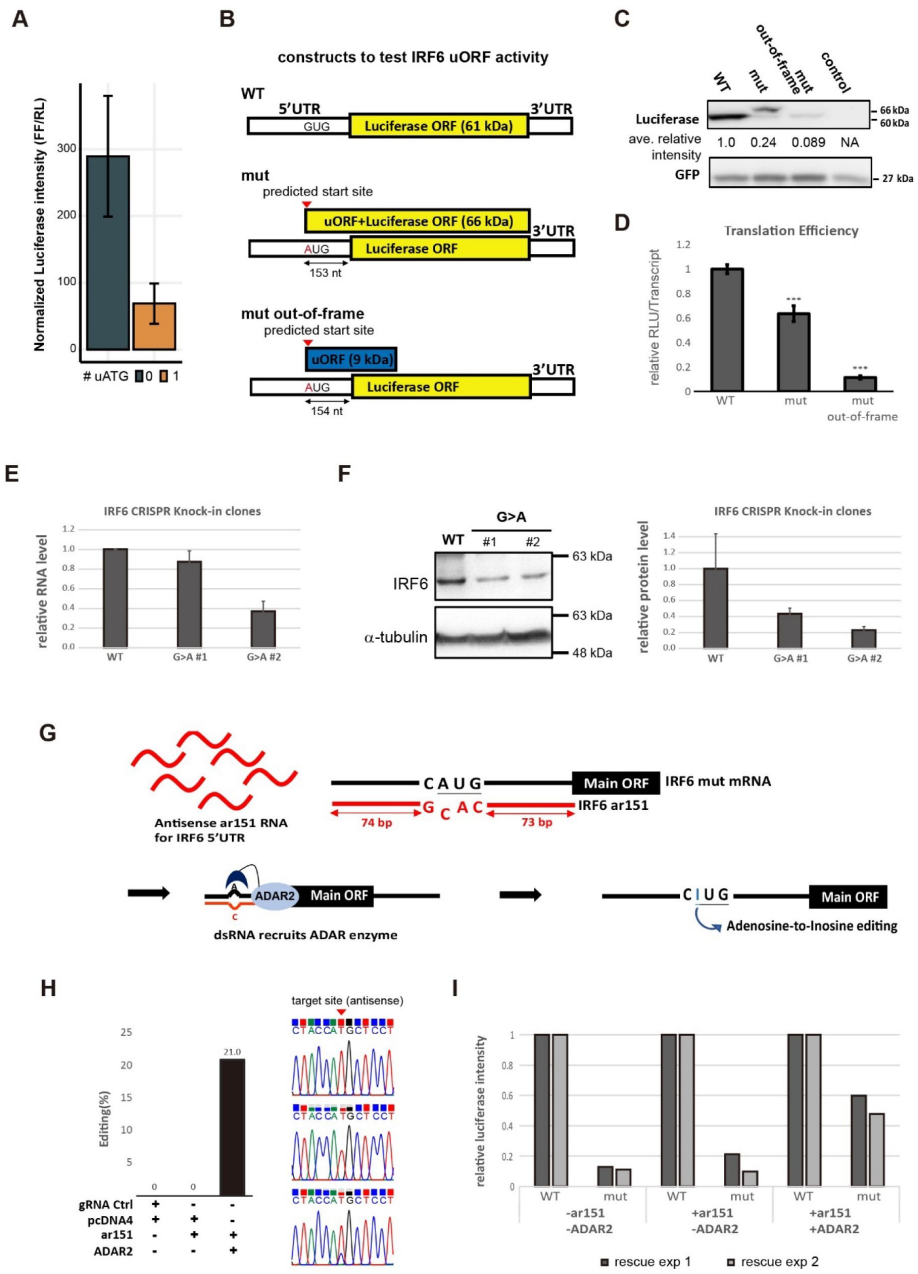


Figure 7.

5' UTR mutations cause translation inhibition by generating upstream ATG

(A) Luciferase activity of eight UTRs with or without a uATG in the 5' UTR. Error bars represent standard errors. **(B)** Constructs of the mutant IRF6 5' UTRs. Mutant (mut) IRF6 generates an in- frame luciferase protein from an upstream ATG, whereas the uORF and luciferase are coded in different frames for the out-of-frame mutant (mut out-of-frame). **(C)** Western blots showing a reduction in luciferase protein levels following 5' UTR mutations. **(D)** Luciferase assays showing a reduction in translation efficiency due to 5' UTR mutations. RLU: relative luminescence units. The individual assays were repeated three times. Statistical significance was determined by a two- tailed Student t-test against WT: ***: $p < 0.001$. **(E)** Relative expression of IRF6 RNA in CRISPR knock-in HEK293T cells. The G>A clones carry the IRF6:c.-4609G>A variant. IRF6 RNA expression was assessed by four independent RT-qPCR experiments and normalized to GAPDH. The error bars represent standard errors. **(F)** IRF6 protein expression in CRISPR knock-in clones. The bar graphs summarize three independent Western blot analyses of IRF6, with normalization to α- tubulin protein levels. The error bars represent standard errors. **(G)** Design of guided ADAR-mediated editing to remove a uORF. **(H)** Editing efficiency achieved by guided ADAR, as determined by Sanger sequencing. **(I)** Restoration of translation efficiency by means of guided ADAR editing, as determined by luciferase assays.

Discussion

In this study, we applied massively parallel polysome assays to examine the effect of mutations in 5' and 3' UTRs in terms of RNA translatability (**Fig. 2**). Our results show that the effect we observed in our high-throughput profiling is reliable, though it could be diluted somewhat in a more complex genomic context (**Fig. 3**). We have shown that T- and GAA- repeat motifs are sensitive to translation-altering mutations (**Fig. 4**), underscoring the involvement of both *cis*-regulatory sequences and *trans*-acting RBPs in regulating RNA translatability. Additionally, we identified the secondary structure of 5' UTRs as a critical factor controlling RNA translatability (**Fig. 5**). We observed that alterations in folding energy, especially in uORF-hosting 5' UTRs, can have detrimental effects, suggesting structure-mediated uORF translation. These findings were summarized by elastic net regression models and robustly explained the polysome-shifting UTR variants (**Fig. 6**). Additionally, the creation of uORFs in 5' UTRs through mutation adversely impacts primary ORF expression (**Fig. 7B-F**), consistent with our finding of pathogenic mutant hotspot at uORF translation start sites (**Fig. 1B**). We also reveal that ADAR-mediated editing can be deployed to rescue primary ORF translation (**Fig. 7G-I**). Our study emphasizes the significant role of precise UTR translation regulation in gene expression and offers a molecular explanation for the impact of UTR mutations on pathogenicity. Below, we discuss some of the strengths and limitations of our study.

Massively parallel evaluation of translation efficiency

Translation efficiency is a key post-transcriptional regulation to control gene output. Efforts have been made to assess translation efficiency systematically in terms of gene structures and sequences (Jia et al., 2020; Leppek et al., 2022; Niederer et al., 2022; Oikonomou et al., 2014; Sample et al., 2019; Savinov et al., 2021; Slutskin et al., 2018; Wissink et al., 2016; Zhao et al., 2014). One way to measure translation efficiency is according to protein output. For instance, in MPRA where reporter sequences are fixed and fused with test sequences, the expression of the reporter fluorescent proteins is determined by Fluorescence Activated Cell Sorting (FACS), with expression being normalized to the level of the other fluorescent reporter from the same construct to account for experimental bias. The tested sequences within each bin are then determined by sequencing. Mass spectrometry can be employed to determine quantitatively endogenous protein levels (Baek et al., 2008). Although this method is relatively quantitative, MPRA has the leverage over RNA stability, especially if the reporter is fused with various UTRs. Moreover, the results for endogenous proteins are influenced by protein turnover, which itself is a very important means of controlling gene expression. Thus, although protein output is a crucial readout of gene expression, this approach does not provide a direct measurement of translation efficiency.

The alternative method to measure translation is through polysome profiling. For endogenous genes, ribosome footprinting can be employed to quantify transcript fragments associated with ribosomes. For MPRA, in which the coding footprint is identical for the whole library, polysome profiling is used instead. Rather than sequencing and then counting transcript fragments attached to ribosomes, polysome profiling does not shear off the transcript, but only fractionates RNAs according to the number (size) of associated ribosomes through ultracentrifugation in a sucrose gradient, representing the most direct way to measure the efficiency of RNAs to recruit ribosomes. However, there are some limitations to this method. First, there is considerable variation in how polysome fractions are used. Some studies have used the average number of ribosomes on each reporter, with each fraction (including the 'ribosome-free' fraction) being considered and the average ribosome number being calculated according to the relative distribution of the transcript in polysomes . In contrast, others use only some of the fractions to infer translation efficiency. For example, percentage translated RNAs in polysomes can be used to infer ribosome loading (Jia et al.,

2020; Schuster et al., 2023), and the monosome to input ratio (recruitment score) can be calculated to infer ribosome recruitment efficiency (Niederer et al., 2022). These alternative approaches to analyzing the data create perplexity to interpret polysome profiles. Another major limitation of polysome profiling is its dependency on promoters. It has been reported that whether polysome profiles reflect translation efficiency is contingent on the promoters driving gene expression (Kong et al., 2008). It remains unclear if the strong expression promoters (such as the CMV promoter) commonly used for MPRA are optimal for revealing differences in polysome profiles. Moreover, because not all ribosomes translate actively on transcripts, they can be slowed, stalled, or collided on a transcript (Meydan & Guydosh, 2021). Although it has been proposed that the major bottleneck of the translation reaction is translation initiation, elongation also contributes to translation efficiency (Riba et al., 2019). Thus, whether the number of ribosomes associated with a transcript truly reflects translation efficiency, especially for endogenous transcripts in diverse coding contexts, remains elusive.

Severe translation defects caused by 5' UTR point mutations

The 5' UTR serves as the major ribosome entry site for translation. In our data, we observed remarkable changes in translation efficiency elicited by 5' UTR point mutations (**Fig. 3**), with IRF6:c.-4609G>A and FTL:c.-161C>T protein outputs being reduced 10-30-fold without affecting respective RNA levels (Supplemental Figs. S3, S4). The impacts of 3' UTR mutations appear to be more additive and context-dependent, consistent with its nature of greater length variation. The most severe UTR mutation we identified in our MPRA dataset is IRF6:c.-4609G>A, which has been observed in Van der Woude syndrome patients. This 5' UTR mutation has been suggested to suppress IRF6 translation by generating a uORF. We have demonstrated herein that IRF6:c.-4609G>A creates a strong upstream translation start codon, dramatically reducing protein output from the primary ORF (**Fig. 7B-F**). We reasoned that a G>A mutation represented an optimal target for ADAR editing, allowing us to switch A back to I, which is read as G by the cellular machinery (**Fig. 7G-I**). Another advantage of ADAR is that editing operates at the RNA level within the potential treatment window, thereby restoring desired protein function without causing any permanent alterations to the DNA.

Ferritin is composed of L (FTL) and H (FTH) subunits, and it is responsible for regulating the storage and intracellular distribution of iron. FTL:c.-161C, one of the UTR variants we observed as eliciting strongly perturbed translation, has been mapped as part of a noncanonical upstream start codon (CAG, TISdb) (Wan & Qian, 2014). It is likely that the C>T mutation disrupts uORF initiation, thereby releasing repression of the primary ORF, and promoting the ~10-fold increase in translation. However, the 5' UTR of FTL is also responsible for forming an IRE stem-loop (Klausner et al., 1993). The FTL:c.-161C>T mutation likely disrupts this hexanucleotide loop structure, preventing it from interacting with the Iron Regulatory Proteins IRP1 and IRP2 (Luscieti et al., 2013). Since IRPs bind to IREs under iron-deficient conditions to suppress translation of both ferritin subunits (Muckenthaler et al., 1998), mutations that abolish IRP-IRE interactions elicit aberrant production of ferritins and are the major causative mutations of Hereditary Hyperferritinaemia Cataract Syndrome. Thus, our massively parallel polysome profiling has uncovered known and novel mutations that cause translation defects, serving as a genetic diagnostic tool for identifying crucial and clinically relevant regulatory mutations.

Data availability

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; *https://www.ncbi.nlm.nih.gov/geo/*) under accession number GSE229492 (reviewer token oxcpuuechhihzuf).

Codes used for the analyses in this study have been deposited at *https://github.com/chienlinglin/modeling-UTR-variants-polysome/*

Other databases used in the study: AREsite2: *http://nibiru.tbi.univie.ac.at/AREsite2* ATtRACT: *https://attract.cnic.es/download*

Acknowledgements

We thank the Genomics Core and the Bioinformatics Core of the Institute of Molecular Biology (IMB), Academia Sinica, for performing the amplicon sequencing and for providing computing resources. We thank all members of IMB for tremendous help and support.

Additional information

Author contributions

C-L L and H-L C designed the experiments. W-P L established the MPRA. Y-L W and J-Y S carried out modeling and analyses. W-P L, Y-C C, H-L C, Y-H C, Y-L K, B-J C, Y-T Hsieh and C-H Y performed experiments. Y-T Huang supervised the statistical analysis. W-P L, Y-C C, Y-L W and C-L L wrote the manuscript.

Funding

This work was supported by a Career Development Award and the Multidisciplinary Health Cloud Research Program of Academia Sinica (AS-CDA-108-M03 and AS-PH-109-01-3), a Career Development Award from the National Health Research Institutes, Taiwan (NHRI-EX112-10908BC), and Excellent Young Scholar Research Grants and a Ta-You Wu Memorial Award from the National Science and Technology Council, Taiwan (NSTC 112-2628-B-001-009, MOST 111-2628-B-001-003 and 108-2118-M-001-013-MY5). H-L C is supported by NSTC Postdoc Fellowship (NSTC 112-2811- B-001-035).

Funding

Academia Sinica (AS-CDA-108-M03)

Academia Sinica (AS-PH-109-01-3)

National Health Research Institutes (NHRI-EX112-10908BC)

National Science and Technology Council (NSTC 112-2628-B-001-009)

National Science and Technology Council (MOST 111-2628-B-001-003)

National Science and Technology Council (MOST 108-2118-M-001-013-MY5)

National Science and Technology Council (NSTC 112-2811-B-001-035)

Additional files

supplemental information

Supplemental Table S1

Supplemental Table S2

Supplemental Table S3

Supplemental Table S4

Supplemental Table S5

Supplemental Table S6

References

1. Abell N. S., DeGorter M. K., Gloudemans M. J., Greenwald E., Smith K. S., He Z. H., Montgomery S. B (2022) **Multiple causal variants underlie genetic associations in humans** *Science* **375**:1247 https://doi.org/10.1126/science.abj5117 | Google Scholar

2. Agarwal V., Bell G. W., Nam J. W., Bartel D. P (2015) **Predicting effective microRNA target sites in mammalian mRNAs** *eLife* **4**:e05005 https://doi.org/10.7554/eLife.05005 | Google Scholar

3. Baek D., Villén J., Shin C., Camargo F. D., Gygi S. P., Bartel D. P (2008) **The impact of microRNAs on protein output** *Nature* **455**:64–U38 https://doi.org/10.1038/nature07242 | Google Scholar

4. Barreau C., Paillard L., Osborne H. B (2005) **AU-rich elements and associated factors: are there unifying principles?** *Nucleic Acids Research* **33**:7138–7150 https://doi.org/10.1093/nar/gki1012 | Google Scholar

5. Beffagna G., Occhi G., Nava A., Vitiello L., Ditadi A., Basso C., Bauce B., Carraro G., Thiene G., Towbin J. A., Danieli G. A., Rampazzo A (2005) **Regulatory mutations in transforming growth factor-β3 gene cause arrhythmogenic right ventricular cardiomyopathy type 1** *Cardiovascular Research* **65**:366–373 https://doi.org/10.1016/j.cardiores.2004.10.005 | Google Scholar

6. Bolger A. M., Lohse M., Usadel B (2014) **Trimmomatic: a flexible trimmer for Illumina sequence data** *Bioinformatics* **30**:2114–2120 https://doi.org/10.1093/bioinformatics/btu170 | Google Scholar

7. Chiang H.-L., Chen Y.-T., Su J.-Y., Lin H.-N., Yu C.-H. A., Hung Y.-J., Wang Y.-L., Huang Y.-T., Lin C.-L (2022) **Mechanism and modeling of human disease- associated near-exon intronic variants that perturb RNA splicing** *Nature Structural & Molecular Biology* https://doi.org/10.1038/s41594-022-00844-1 | Google Scholar

8. Choe J., Lin S. B., Zhang W. C., Liu Q., Wang L. F., Ramirez-Moya J., Du P., Kim W., Tang S. J., Sliz P., Santisteban P., George R. E., Richards W. G., Wong K. K., Locker N., Slack F. J., Gregory R. I (2018) **mRNA circularization by METTL3-eIF3h enhances translation and promotes oncogenesis** *Nature* **561**:556 https://doi.org/10.1038/s41586-018-0538-8 | Google Scholar

9. Danecek P., Bonfield J. K., Liddle J., Marshall J., Ohan V., Pollard M. O., Whitwham A., Keane T., McCarthy S. A., Davies R. M., Li H (2021) **Twelve years of SAMtools and BCFtools** *Gigascience* **10**:giab008 https://doi.org/10.1093/gigascience/giab008 | Google Scholar

10. de Lima R. L. L. F., Hoper S. A., Ghassibe M., Cooper M. E., Rorick N. K., Kondo S., Katz L., Marazita M. L., Compton J., Bale S., Hehr U., Dixon M. J., Daack- Hirsch S., Boute O., Bayet B., Revencu N., Verellen-Dumoulin C., Vikkula M., Richieri-Costa A., Schutte . . ., C B. (2009) **Prevalence and nonrandom distribution of exonic mutations in interferon regulatory factor 6 in 307 families with Van der Woude syndrome and 37 families with popliteal pterygium syndrome.** *Genetics in Medicine* **11**:241–247 https://doi.org/10.1097/GIM.0b013e318197a49a | Google Scholar

11. Dusl M., Senderek J., Müller J. S., Vogel J. G., Pertl A., Stucka R., Lochmüller H., David R., Abicht A (2015) **A 3′-UTR mutation creates a microRNA target site in the gene of patients with congenital myasthenic syndrome** *Human Molecular Genetics* **24**:3418–3426 https://doi.org/10.1093/hmg/ddv090 | Google Scholar

12. Fay M. M., Lyons S. M., Ivanov P (2017) **RNA G-Quadruplexes in Biology: Principles and Molecular Mechanisms** *Journal of Molecular Biology* **429**:2127–2147 https://doi.org/10.1016/j.jmb.2017.05.017 | Google Scholar

13. Gandin V., Sikstrom K., Alain T., Morita M., McLaughlan S., Larsson O., Topisirovic I (2014) **Polysome fractionation and analysis of mammalian translatomes on a genome-wide scale** *J Vis Exp* **87** https://doi.org/10.3791/51455 | Google Scholar

14. Georgakopoulos-Soares I., Parada G. E., Hemberg M (2022) **Secondary structures in RNA synthesis, splicing and translation** *Computational and Structural Biotechnology Journal* **20**:2871–2884 https://doi.org/10.1016/j.csbj.2022.05.041 | Google Scholar

15. Giudice G., Sanchez-Cabo F., Torroja C., Lara-Pezzi E (2016) **ATtRACT-a database of RNA-binding proteins and associated motifs** *Database-the Journal of Biological Databases and Curation* :baw035 https://doi.org/10.1093/database/baw035 | Google Scholar

16. Griesemer D., Xue J. R., Reilly S. K., Ulirsch J. C., Kukreja K., Davis J. R., Kanai M., Yang D. K., Butts J. C., Guney M. H., Luban J., Montgomery S. B., Finucane H. K., Novina C. D., Tewhey R., Sabeti P. C (2021) **Genome-wide functional screen of 3 ′ UTR variants uncovers causal variants for human disease and evolution** *Cell* **184**:5247 https://doi.org/10.1016/j.cell.2021.08.025 | Google Scholar

17. Guenther U. P., Weinberg D. E., Zubradt M. M., Tedeschi F. A., Stawicki B. N., Zagore L. L., Brar G. A., Licatalosi D. D., Bartel D. P., Weissman J. S., Jankowsky E (2018) **The helicase Ded1p controls use of near-cognate translation initiation codons in 5 ' UTRs** *Nature* **559**:130 https://doi.org/10.1038/s41586-018-0258-0 | Google Scholar

18. Jia L. F., Mao Y. H., Ji Q. Q., Dersh D., Yewdell J. W., Qian S. B (2020) **Decoding mRNA translatability and stability from the 5 ′ UTR** *Nature Structural & Molecular Biology* **27**:814 https://doi.org/10.1038/s41594-020-0465-x | Google Scholar

19. Kerpedjiev P., Hammer S., Hofacker I. L (2015) **Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams** *Bioinformatics* **31**:3374–3376 https://doi.org/10.1093/bioinformatics/btv372 | Google Scholar

20. Klausner R. D., Rouault T. A., Harford J. B (1993) **Regulating the Fate of Messenger- Rna - the Control of Cellular Iron-Metabolism** *Cell* **72**:19–28 https://doi.org/10.1016/0092-8674(93)90046-S | Google Scholar

21. Komar A. A., Hatzoglou M (2005) **Internal ribosome entry sites in cellular mRNAs: Mystery of their existence** *Journal of Biological Chemistry* **280**:23425–23428 https://doi.org/10.1074/jbc.R400041200 | Google Scholar

22. Kong Y. W., Cannell I. G., de Moor C. H., Hill K., Garside P. G., Hamilton T. L., Meijer H. A., Dobbyn H. C., Stoneley M., Spriggs K. A., Willis A. E., Bushell M. (2008) **The mechanism of micro-RNA-mediated translation repression is determined by the promoter of the target**

gene *Proceedings of the National Academy of Sciences of the United States of America* **105**:8866–8871 https://doi.org/10.1073/pnas.0800650105 | Google Scholar

23. Kuhn M (2008) **Building Predictive Models in R Using the caret Package** *Journal of Statistical Software* **28**:1–26 https://doi.org/10.18637/jss.v028.i05 | Google Scholar

24. Kumari S., Bugaut A., Huppert J. L., Balasubramanian S (2007) **An RNA G- quadruplex in the 5 ' UTR of the NRAS proto-oncogene modulates translation** *Nature Chemical Biology* **3**:218–221 https://doi.org/10.1038/nchembio864 | Google Scholar

25. Landrum M. J., Lee J. M., Benson M., Brown G., Chao C., Chitipiralla S., Gu B., Hart J., Hoffman D., Hoover J., Jang W., Katz K., Ovetsky M., Riley G., Sethi A., Tully R., Villamarin-Salomon R., Rubinstein W., Maglott D. R (2016) **ClinVar: public archive of interpretations of clinically relevant variants** *Nucleic Acids Research* **44**:D862–D868 https://doi.org/10.1093/nar/gkv1222 | Google Scholar

26. Leppek K., Byeon G. W., Kladwang W., Wayment-Steele H. K., Kerr C. H., Xu A. D. F., Kim D., Topkar V. V., Choe C., Rothschild D., Tiu G. C., Wellington-Oguri R., Fujii K., Sharma E., Watkins A. M., Nicol J. J., Romano J., Tunguz B., Diaz F., ..., Das R. (2022) **Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics** *Nature Communications* **13** https://doi.org/10.1038/s41467-022-28776-w | Google Scholar

27. Leppek K., Das R., Barna M (2018) **Functional 5 ' UTR mRNA structures in eukaryotic translation regulation and how to find them** *Nature Reviews Molecular Cell Biology* **19**:673–673 https://doi.org/10.1038/s41580-018-0055-5 | Google Scholar

28. Lim Y. T., Arora S., Schuster S. L., Corey L., Fitzgibbon M., Wladyka C. L., Wu X. Y., Coleman I. M., Delrow J. J., Corey E., True L. D., Nelson P. S., Ha G., Hsieh A. C (2021) **Multiplexed functional genomic analysis of 5' untranslated region mutations across the spectrum of prostate cancer** *Nature Communications* **12** https://doi.org/10.1038/s41467-021-24445-6 | Google Scholar

29. Litterman A. J., Kageyama R., Le Tonqueze O., Zhao W. X., Gagnon J. D., Goodarzi H., Erle D. J., Ansel K. M. (2019) **A massively parallel 3 ' UTR reporter assay reveals relationships between nucleotide content, sequence conservation, and mRNA destabilization** *Genome Research* **29**:896–906 https://doi.org/10.1101/gr.242552.118 | Google Scholar

30. Lorenz R., Bernhart S. H., Siederdissen C. H. Z., Tafer H., Flamm C., Stadler P. F., Hofacker I. L (2011) **ViennaRNA Package 2.0** *Algorithms for Molecular Biology* **6** https://doi.org/10.1186/1748-7188-6-26 | Google Scholar

31. Luscieti S., Tolle G., Aranda J., Benet-Campos C., Risse F., Moran E., Muckenthaler M. U., Sanchez M (2013) **Novel Mutations in the Ferritin-L Iron-Responsive Element That Only Mildly Impair Irp Binding Cause Hereditary Hyperferritinaemia Cataract Syndrome** *American Journal of Hematology* **88**:E63–E63 Google Scholar

32. MacArthur J., Bowler E., Cerezo M., Gil L., Hall P., Hastings E., Junkins H., McMahon A., Milano A., Morales J., Pendlington Z. M., Welter D., Burdett T., Hindorff L., Flicek P., Cunningham F., Parkinson H (2017) **The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)** *Nucleic Acids Research* **45**:D896–D901 https://doi.org/10.1093/nar/gkw1133 | Google Scholar

33. Mayya V. K., Duchaine T. F (2019) **Ciphers and Executioners: How 3 '-Untranslated Regions Determine the Fate of Messenger RNAs** *Frontiers in Genetics* **10** https://doi.org/10.3389

/fgene.2019.00006 | Google Scholar

34. Mazumder B., Seshadri V., Fox P. L (2003) **Translational control by the 3 ' -UTR:: the ends specify the means** *Trends in Biochemical Sciences* **28**:91–98 https://doi.org/10.1016/S0968-0004(03)00002-1 | Google Scholar

35. Merkle T., Merz S., Reautschnig P., Blaha A., Li Q., Vogel P., Wettengel J., Li J. B., Stafforst T (2019) **Precise RNA editing by recruiting endogenous ADARs with antisense oligonucleotides** *Nature Biotechnology* **37**:133 https://doi.org/10.1038/s41587-019-0013-6 | Google Scholar

36. Meydan S., Guydosh N. R (2021) **A cellular handbook for collided ribosomes: surveillance pathways and collision types** *Current Genetics* **67**:19–26 https://doi.org/10.1007/s00294-020-01111-w | Google Scholar

37. Mikl M., Eletto D., Nijim M., Lee M., Lafzi A., Mhamedi F., David O., Sain S. B., Handler K., Moor A. E (2022) **A massively parallel reporter assay reveals focused and broadly encoded RNA localization signals in neurons** *Nucleic Acids Research* **50**:10643–10664 https://doi.org/10.1093/nar/gkac806 | Google Scholar

38. Muckenthaler M., Gray N. K., Hentze M. W (1998) **IRP-1 binding to ferritin mRNA prevents the recruitment of the small ribosomal subunit by the cap-binding complex eIF4F** *Molecular Cell* **2**:383–388 https://doi.org/10.1016/S1097-2765(00)80282-8 | Google Scholar

39. Niederer R. O., Rojas-Duran M. F., Zinshteyn B., Gilbert W. V (2022) **Direct analysis of ribosome targeting illuminates thousand-fold regulation of translation initiation** *Cell Systems* **13**:256 https://doi.org/10.1016/j.cels.2021.12.002 | Google Scholar

40. Oikonomou P., Goodarzi H., Tavazoie S (2014) **Systematic Identification of Regulatory Elements in Conserved 3 ' UTRs of Human Transcripts** *Cell Reports* **7**:281–292 https://doi.org/10.1016/j.celrep.2014.03.001 | Google Scholar

41. Pagès H A. P., Gentleman R, DebRoy S. (2022) **Biostrings: Efficient manipulation of biological strings** *R package* v2.66.0 https://bioconductor.org/packages/Biostrings

42. Qu L., Yi Z. Y., Zhu S. Y., Wang C. H., Cao Z. Z., Zhou Z., Yuan P. F., Yu Y., Tian F., Liu Z. H., Bao Y., Zhao Y. X., Wei W. S (2019) **Programmable RNA editing by recruiting endogenous ADAR using engineered RNAs** *Nature Biotechnology* **37**:1380–1380 https://doi.org/10.1038/s41587-019-0292-y | Google Scholar

43. Quinlan A. R., Hall I. M (2010) **BEDTools: a flexible suite of utilities for comparing genomic features** *Bioinformatics* **26**:841–842 https://doi.org/10.1093/bioinformatics/btq033 | Google Scholar

44. R-Core-Team (2021) **R: A Language and Environment for Statistical Computing** Vienna, Austria: R Foundation for Statistical Computing https://www.R-project.org

45. Rabani M., Pieper L., Chew G. L., Schier A. F (2017) **A Massively Parallel Reporter Assay of 3 ' UTR Sequences Identifies In Vivo Rules for mRNA Degradation** *Molecular Cell* **68**:1083 https://doi.org/10.1016/j.molcel.2017.11.014 | Google Scholar

46. Riba A., Di Nanni N., Mittal N., Arhné E., Schmidt A., Zavolan M. (2019) **Protein synthesis rates and ribosome occupancies reveal determinants of translation elongation rates**

Proceedings of the National Academy of Sciences of the United States of America **116**:15023–15032 https://doi.org/10.1073/pnas.1817299116 | Google Scholar

47. Saito T., Rehmsmeier M (2017) **Precrec: fast and accurate precision-recall and ROC curve calculations in R** *Bioinformatics* **33**:145–147 https://doi.org/10.1093/bioinformatics/btw570 | Google Scholar

48. Sample P. J., Wang B., Reid D. W., Presnyak V., McFadyen I. J., Morris D. R., Seelig G (2019) **Human 5 ′ UTR design and variant effect prediction from a massively parallel translation assay** *Nature Biotechnology* **37**:803 https://doi.org/10.1038/s41587-019-0164-5 | Google Scholar

49. Savinov A., Brandsen B. M., Angell B. E., Cuperus J. T., Fields S (2021) **Effects of sequence motifs in the yeast 3 ′ untranslated region determined from massively parallel assays of random sequences** *Genome Biology* **22** https://doi.org/10.1186/s13059-021-02509-6 | Google Scholar

50. Schuster S. L., Arora S., Wladyka C. L., Itagi P., Corey L., Young D., Stackhouse B. L., Kollath L., Wu Q. V., Corey E., True L. D., Ha G., Paddison P. J., Hsieh A. C (2023) **Multi-level functional genomics reveals molecular and cellular oncogenicity of patient-based 3′ untranslated region mutations** *Cell Reports* **42**:112840 https://doi.org/10.1016/j.celrep.2023.112840 | Google Scholar

51. Serikawa T., Spanos C., von Hacht A., Budisa N., Rappsilber J., Kurreck J. (2018) **Comprehensive identification of proteins binding to RNA G-quadruplex motifs in the 5 ' UTR of tumor-associated mRNAs** *Biochimie* **144**:169–184 https://doi.org/10.1016/j.biochi.2017.11.003 | Google Scholar

52. Sherry S. T., Ward M. H., Sirotkin K (1999) **dbSNP - Database for single nucleotide polymorphisms and other classes of minor genetic variation** *Genome Research* **9**:677–679 Google Scholar

53. Sliskovic I., Eich H., Muller-McNicoll M (2022) **Exploring the multifunctionality of SR proteins** *Biochem Soc Trans* **50**:187–198 https://doi.org/10.1042/BST20210325 | Google Scholar

54. Slutskin I. V., Weingarten-Gabbay S., Nir R., Weinberger A., Segal E (2018) **Unraveling the determinants of microRNA mediated regulation using a massively parallel reporter assay** *Nature Communications* **9** https://doi.org/10.1038/s41467-018-02980-z | Google Scholar

55. Smith T., Heger A., Sudbery I (2017) **UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy** *Genome Research* **27**:491–499 https://doi.org/10.1101/gr.209601.116 | Google Scholar

56. Stenson P. D., Mort M., Ball E. V., Evans K., Hayden M., Heywood S., Hussain M., Phillips A. D., Cooper D. N (2017) **The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies** *Human Genetics* **136**:665–677 https://doi.org/10.1007/s00439-017-1779-6 | Google Scholar

57. Steri M., Idda M. L., Whalen M. B., Orru V (2018) **Genetic variants in mRNA untranslated regions** *Wiley Interdisciplinary Reviews-Rna* **9**:e1474 https://doi.org/10.1002/wrna.1474 | Google Scholar

58. Stoecklin G., Stoeckle P., Lu M., Muehlemann O., Moroni C (2001) **Cellular mutants define a common mRNA degradation pathway targeting cytokine AU-rich elements** *Rna* **7**:1578–1588 Google Scholar

59. Thul P. J., Åkesson L., Wiking M., Mahdessian D., Geladaki A., Blal H. A., Alm T., Asplund A., Björk L., Breckels L. M., Bäckström A., Danielsson F., Fagerberg L., Fall J., Gatto L., Gnann C., Hober S., Hjelmare M., Johansson F., ..., Lundberg E. (2017) **A subcellular map of the human proteome** *Science* **356**:eaal3321 https://doi.org/10.1126/science.aal3321 | Google Scholar

60. Van Etten J., Schagat T. L., Hrit J., Weidmann C. A., Brumbaugh J., Coon J. J., Goldstrohm A. C. (2012) **Human Pumilio Proteins Recruit Multiple Deadenylases to Efficiently Repress Messenger RNAs** *Journal of Biological Chemistry* **287**:36370–36383 https://doi.org/10.1074/jbc.M112.373522 | Google Scholar

61. Wan J., Qian S. B (2014) **TISdb: a database for alternative translation initiation in mammalian cells** *Nucleic Acids Research* **42**:D845–D850 https://doi.org/10.1093/nar/gkt1085 | Google Scholar

62. Wells S. E., Hillner P. E., Vale R. D., Sachs A. B (1998) **Circularization of mRNA by eukaryotic translation initiation factors** *Molecular Cell* **2**:135–140 https://doi.org/10.1016/S1097-2765(00)80122-7 | Google Scholar

63. Wissink E. M., Fogarty E. A., Grimson A (2016) **High-throughput discovery of post-transcriptional cis-regulatory elements** *Bmc Genomics* **17** https://doi.org/10.1186/s12864-016-2479-7 | Google Scholar

64. Xiang Y. Z., Huang W. Z., Tan L. M., Chen T. Y., He Y., Irving P. S., Weeks K. M., Zhang Q. C., Dong X. N (2023) **Pervasive downstream RNA hairpins dynamically dictate start-codon selection** *Nature* https://doi.org/10.1038/s41586-023-06500 | Google Scholar

65. Zhao W. X., Pollack J. L., Blagev D. P., Zaitlen N., McManus M. T., Erle D. J (2014) **Massively parallel functional annotation of 3 ' untranslated regions** *Nature Biotechnology* **32**:387–U259 https://doi.org/10.1038/nbt.2851 | Google Scholar

Author information

Wei-Ping Li[†]
Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan, Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

[†]These authors contributed equally

Jia-Ying Su[†]
Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan, Institute of Statistical Science, Academia Sinica, Taipei, Taiwan
ORCID iD: 0000-0001-5934-5458

[†]These authors contributed equally

Yu-Chi Chang
Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan

Hung-Lun Chiang
Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan

Yun-Lin Wang
Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan

Ang-Chu Huang
Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan

Yu-Tung Hsieh
Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan

Yi-Hsuan Chiang
Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan

Yen-Ling Ko
Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan

Bing-Jen Chiang
Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan

Cheng-Han Yang
Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan

Yen-Tsung Huang
Institute of Statistical Science, Academia Sinica, Taipei, Taiwan
ORCID iD: 0000-0001-7657-0040

Chien-Ling Lin
Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan
ORCID iD: 0000-0002-5730-799X

For correspondence: mbcllin@gate.sinica.edu.tw

Editors

Reviewing Editor
Murim Choi
Seoul National University, Seoul, Republic of Korea

Senior Editor
Murim Choi
Seoul National University, Seoul, Republic of Korea

Reviewer #1 (Public review):

The authors describe a massively parallel reporter assays (MPRA) screen focused at identifying polymorphisms in 5' and 3' UTRs that affect translation efficiency and thus might

have a functional impact on cells. The topic is of timely interest, and indeed, several related efforts have recently been published and preprinted (e.g., https://pubmed.ncbi.nlm.nih.gov/37516102/ and https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10635273/). This study has several major issues with the results and their presentation.

Major comments:

• The main issue remains that it appears that the screen has largely failed, and the reasons for that remain unclear, which make it difficult to interpret how useful is the resulting data. The authors mention batch effects as a potential contributor. The authors start with a library that includes ~6,000 variants, which makes it a medium-size MPRA. But then, only 483 pairs of WT/mutated UTRs yield high confidence information, which is already a small number for any downstream statistical analysis, particularly since most don't actually affect translation in the reporter screen setting (which is not unexpected). It is unclear why >90% of the library did not give high-confidence information. The profiles presented as base-case examples in Fig. 2B don't look very informative or convincing. All the subsequent analysis is done on a very small set of UTRs that have an effect, and it is unclear to this reviewer how these can yield statistically significant and/or biologically-relevant associations.

• From the variants that had an effect, the authors go on to carry out some protein-level validations, and see some changes, but it is not clear if those changes are in the same direction was observed in the screen. In their rebuttal the authors explain that they largely can not infer directionality of changes form the screen, which further limits its utility.

• It is particularly puzzling how the authors can build a machine learning predictor with >3,000 features when the dataset they use for training the model has just a few dozens of translation-shifting variants.

Comments on revisions:

It appears that the authors have extracted the information they could from the problematic dataset they obtained. Repeating the experiments in a cleaner setting, obtaining data for the >6000 UTRs they intended will allow the authors to achieve the goals they set out to achieve in establishing the screen.

https://doi.org/10.7554/eLife.98814.2.sa1

Author response:

The following is the authors' response to the current reviews.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public review):***
>
> *The authors describe a massively parallel reporter assays (MPRA) screen focused at identifying polymorphisms in 5' and 3' UTRs that affect translation efficiency and thus might have a functional impact on cells. The topic is of timely interest, and indeed, several related efforts have recently been published and preprinted (e.g., https://pubmed.ncbi.nlm.nih.gov/37516102/ and https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10635273/). This study has several major issues with the results and their presentation.*
>
> *Major comments:*
>
> *• The main issue remains that it appears that the screen has largely failed, and the reasons for that remain unclear, which make it difficult to interpret how useful is the*

resulting data. The authors mention batch effects as a potential contributor. The authors start with a library that includes ~6,000 variants, which makes it a medium-size MPRA. But then, only 483 pairs of WT/mutated UTRs yield high confidence information, which is already a small number for any downstream statistical analysis, particularly since most don't actually affect translation in the reporter screen setting (which is not unexpected). It is unclear why >90% of the library did not give high-confidence information. The profiles presented as base-case examples in Fig. 2B don't look very informative or convincing. All the subsequent analysis is done on a very small set of UTRs that have an effect, and it is unclear to this reviewer how these can yield statistically significant and/or biologically-relevant associations.

• From the variants that had an effect, the authors go on to carry out some protein-level validations, and see some changes, but it is not clear if those changes are in the same direction was observed in the screen. In their rebuttal the authors explain that they largely can not infer directionality of changes form the screen, which further limits its utility.

• It is particularly puzzling how the authors can build a machine learning predictor with >3,000 features when the dataset they use for training the model has just a few dozens of translation-shifting variants.

We recognize that RNA distribution within polysomes is inherently less stable than the associated protein components. This instability has been noted in previous studies, including those cited by the reviewer, which used RNA from bulk polysomes to infer the translatome without fractionation. Acknowledging this limitation, we purposely adopted a conservative strategy: (i) performing gross fractionation of polysomes, and (ii) collaborating with biostatisticians at the Institute of Statistical Science, Academia Sinica, to design a conservative yet optimized analysis pipeline that minimized batch effects.

This approach proved robust: representative cases in Fig. 2B clearly demonstrate distinct distributions of reference and alternative alleles. From our high-confidence dataset, we applied a well-established statistical framework specifically designed to accommodate multiple influencing factors in relatively small datasets (Elements of Statistical Learning by Hastie, Tibshirani, and Friedman). We further conducted sensitivity analyses to select an optimal QC cutoff across a range of stringencies, ensuring maximal reliability of our results. We have therefore successfully shortlisted UTR variants which have strong effect on translation.

Building upon these conservative measures, we developed a predictive model for translation effects of UTR variants. Importantly, this model was validated not only with our internal test dataset but also with independent external datasets. In addition, the sequence features identified by the model were validated through reporter assays and *in vivo* CRISPR editing. These external and functional validations establish the generalizability and robustness of our approach.

A more detailed analysis of the directionality of changes in translation efficiency is under active investigation. These results will be reported in a separate manuscript currently in preparation.

The following is the authors' response to the original reviews.

Public Reviews:

> ***Reviewer #1 (Public Review):***
>
> *The authors describe a massively parallel reporter assays (MPRA) screen focused on identifying polymorphisms in 5' and 3' UTRs that affect translation efficiency and thus might have a functional impact on cells. The topic is of timely interest, and indeed, several related efforts have recently been published and preprinted (e.g., https://pubmed.ncbi.nlm.nih.gov/37516102/ and https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10635273/). This study has several major issues with the results and their presentation.*
>
> *Major comments:*
>
> *(1) The main issue is that it appears that the screen has largely failed, yet the reasons for that are unclear, which makes it difficult to interpret. The authors start with a library that includes approximately 6,000 variants, which makes it a medium-sized MPRA. But then, only 483 pairs of WT/mutated UTRs yield highconfidence information, which is already a small number for any downstream statistical analysis, particularly since most don't actually affect translation in the reporter screen setting (which is not unexpected). It is unclear why >90% of the library did not give high-confidence information. The profiles presented as basecase examples in Figure 2B don't look very informative or convincing. All the subsequent analysis is done on a very small set of UTRs that have an effect, and it is unclear to this reviewer how these can yield statistically significant and/or biologically relevant associations.*

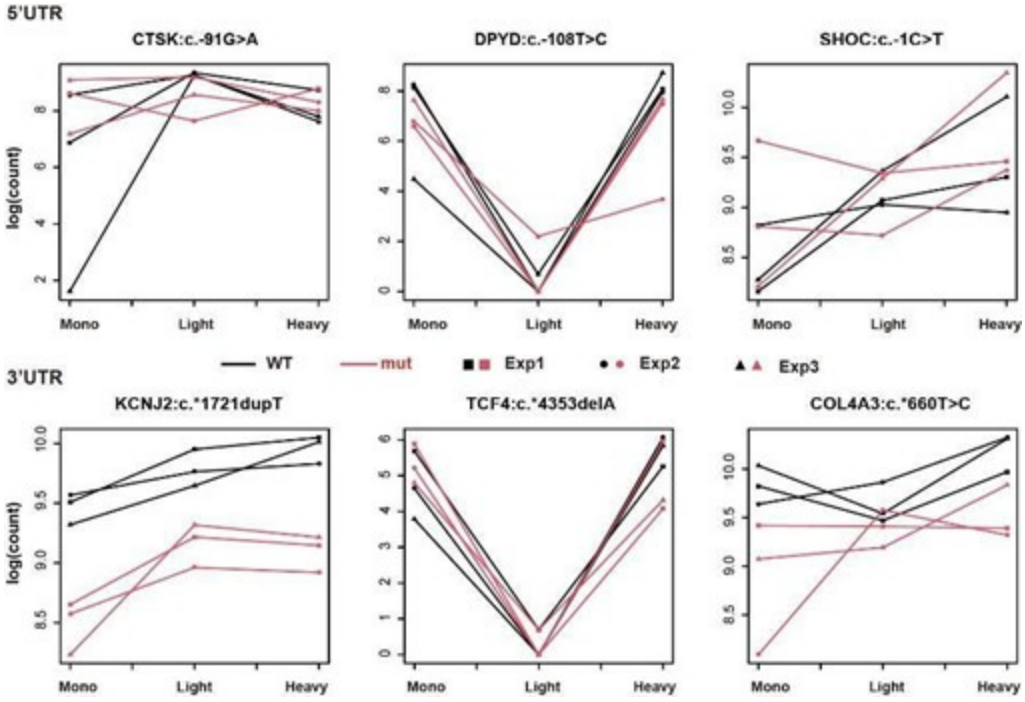
To make sure our final results are technically and statistically sound, we applied stringent selection criteria and cutoffs in our analytics workflow. First, from our RNA-seq dataset, we filtered the UTRs with at least 20 reads in a polysome profile across all three repeated experiments. Secondly, in the following main analysis using a negative binomial generalized linear model (GLM), we further excluded the UTRs that displayed batch effect, i.e. their batch-related main effect and interaction are significant. We believe our measure has safeguarded the filtered observations (UTRs) from the (potential) high variation of our massively parallel translation assays and thus gives high confidence to our results.

Regarding the interpretation of Figure 2B, since we aimed to identify the UTRs whose interaction term of genotype and fractions is significant in our generalized linear model, it is statistically conventional to doublecheck the interaction of the two variables using such a graph. For instance, in the top left panel of Figure 2B (5'UTR of ANK2:c.-39G>T), we can see that read counts of WT samples congruously decreased from Mono to Light, whereas the read counts of mutant samples were roughly the same in the two fractions – the trend is different between WT and mutant. Ergo, the distinct distribution patterns of two genotypes across three fractions in Figure 2B offer the readers a convincing visual supplement to our statistics from GLM.

In contrast to Figure 2B, the graphs of nonsignificant UTRs (shown below) reveal that the trends between the two genotypes are similar across the 'Mono and Light' and 'Light and Heavy' polysome fractions. Importantly, our analysis remains unaffected by differential expression levels between WT and mutant, as it specifically distinguishes polysome profiles with different distributions. This consistent trend further supports the lack of interaction between genotype and polysome fractions for these UTRs.

Author response image 1.

Examples of non-significant UTR pairs in massively parallel polysome profiling assays.



> *(2) From the variants that had an effect, the authors go on to carry out some protein-level validations and see some changes, but it is not clear if those changes are in the same direction as observed in the screen.*

To infer the directionality of translation efficiency from polysome profiles, a common approach involves pooling polysome fractions and comparing them with free or monosome fractions to identify 'translating' fractions. However, this method has two major potential pitfalls: (i) it sacrifices resolution and does not account for potential bias toward light or heavy polysomes, and (ii) it fails to account for discrepancies between polysome load and actual protein output (as discussed in https://doi.org/10.1016/j.celrep.2024.114098 and https://doi.org/10.1038/s41598-019-47424-w). Therefore, our analysis focused on the changes within polysome profiles themselves. 'Significant' candidates were identified based on a significant interaction between genotype and polysome distribution using a negative binomial generalized linear model, without presupposing the direction of change on protein output.

> *(3) The authors follow up on specific motifs and specific RBPs predicted to bind them, but it is unclear how many of the hits in the screen actually have these motifs, or how significant motifs can arise from such a small sample size.*

We calculated the Δmotif enrichment in significant UTRs versus nonsignificant UTRs using Fisher's exact test. For example, the enrichment of the Δ'AGGG' motif in 3' UTRs is shown below:

Author response table 1.

Motif (AGGG)	HC-Significant 3' UTRs	HC-Nonsignificant 3' UTRs
Change of motif	3	1
No change of motif	24	223

This test yields a *P-value* of 0.004167 by Fisher's exact test. The *P-values* and Odds ratios of Δmotifs in relation to polysome shifting are included in Supplementary Table S4, and we will update the detailed motif information in the revised Supplementary Table S4.

> *(4) It is particularly puzzling how the authors can build a machine learning predictor with >3,000 features when the dataset they use for training the model has just a few dozens of translation-shifting variants.*

We understand the concern regarding the relatively small number of translation-shifting variants compared to the large number of features. To address this, we employed LASSO regression, which, according to The Elements of Statistical Learning by Hastie, Tibshirani, and Friedman, is particularly suitable for datasets where the number of features pp is much larger than the number of samples NN. LASSO effectively performs feature selection by shrinking less important coefficients to zero, allowing us to build a robust and generalizable model despite the limited number of variants.

> *(5) The lack of meaningful validation experiments altering the SNPs in the endogenous loci by genome editing limits the impact of the results.*

Following the reviewer's suggestion, we assessed the endogenous mutant effect by generating CRISPR knock-in clones carrying the IRF6:c.-4609G>A variant. We showed that this G>A variant generate a deleterious upstream open reading frame, which dramatically reduced protein expression of the main open reading frame (Fig. 7B-D). The genome editing further demonstrated the G>A variant reduced endogenous IRF6 protein expression to 23% or 44% in two independent clones. We have incorporated the genome editing results in the revised main text and the new Figure 7E&F:

"To further validate the endogenous effect of the novel upstream ATG (uATG), we generated CRISPR knockin clones carrying the IRF6:c.-4609G>A variant and examined its impact on gene expression. The introduction of the uATG reduced RNA levels to 88% and 37% of the wild-type in two independent clones (Fig. 7E), and protein levels to 44% and 23%, respectively (Fig. 7F), resulting in an overall reduction of translation efficiency to 50–62%." (p.18)

> ***Reviewer #2 (Public Review):***
>
> *Summary:*
>
> *In their paper "Massively Parallel Polyribosome Profiling Reveals Translation Defects of Human DiseaseRelevant UTR Mutations" the authors use massively parallel polysome profiling to determine the effects of 5' and 3' UTR SNPs (from dbSNP/ClinVar) on translational output. They show that some UTR SNPs cause a change in the polysome profile with respect to the wild-type and that pathogenic SNPs are enriched in the polysome-shifting group. They validate that some changes in polysome profiles are predictive of differences in translational output using transiently expressed luciferase reporters. Additionally, they identify sequence motifs enriched in the polysome-shifting group. They show that 2 enriched 5' UTR motifs increase the translation of a luciferase reporter in a protein-dependent manner, highlighting the use of their method to identify translational control elements.*

Strengths:

This is a useful method and approach, as UTR variants have been more difficult to study than coding variants. Additionally, their evidence that pathogenic mutations are more likely to cause changes in polysome association is well supported.

> *Weaknesses:*
>
> *The authors acknowledge that they "did not intend to immediately translate the altered polysome profile into an increase or decrease in translation efficiency, as the direction of the shift was not readily evident. Additionally, sedimentation in the sucrose gradient may have been partially affected by heavy particles other than ribosomes." However, shifted polysome distribution is used as a category for many downstream analyses. Without further clarity or subdivision, it is very difficult to interpret the results (for example in Figure 5A, is it surprising that the polysome shifting mutants decrease structure? Are the polysome "shifts" towards the untranslated or heavy fractions?)*

Our approach, combining polysome fractionation of the UTR library with negative binomial generalized linear model (GLM) analysis of RNA-seq data, systematically identifies variants that affect translational efficiency. The GLM model is specifically designed to detect UTR pairs with significant interactions between genotype and polysome fractions, relying solely on changes in polysome profiles to identify variants that disrupt translation. Consequently, our analytical method does not determine the direction of translation alteration.

Following the massively parallel polysome profiling, we sought to understand how these polysome-shifting variants influence the translation process. To do this, we examined their effects on RNA characteristics related to translation, such as RBP binding and RNA structure. In Figure 5A, we observed a notable trend in significant hits within 5' UTRs—they tend to increase ΔG (weaker folding energy) in response to changes in polysome profiles, regardless of whether protein production increases or decreases (Fig. 3).

> ***Recommendations for the authors:***
>
> ***Reviewer #1 (Recommendations For The Authors):***
>
> *Minor comments:*
>
> *(1) Figure 3A - the claim that 5'UTR variants had a stronger effect than 3'UTR is based on the two UTRs with the strongest effect. It is unclear how these differences between 5' and 3'UTRs are significant.*

We carried out a Wilcoxon rank-sum test to examine the mut/WT fold change of translation efficiency between the 3' and 5' UTR variants. The results showed that the 5' UTR variants exhibited a greater change of translation efficiency. We have inserted this result in the revised Figure 3C and refers to this figure in the main text: "Furthermore, we observed that 5' UTR variants had a greater impact on translation activity relative to 3' UTR variants (Fig. 3C)." (p. 12)

> *(2) Figures 2B and S1, S2 - what is the meaning of less signal for a light chain and a similar signal for a heavy chain? How can this situation, while being a significant difference between the profiles, lead to a biologically relevant difference in eventual protein output?*

Taking 3'UTR ACADSB:c.*4177G>A (bottom-left panel in Figure 2B) as an example: WT transcripts have less read count (in the unit of log(CPM)) compared with the transcripts carrying the mutant UTR in the light polysome-containing fraction, whereas the read counts of the two genotypes are approximately the same in the heavy polysome-containing fraction.

In line with our reply to Reviewer 1's major comment 1, we aimed to identify the UTRs whose interaction term of genotype and fractions is significant in our generalized linear model (GLM). That is, the UTR pairs whose WT and mutant have different trends across the fractions

(Mono to Light & Light to Heavy) are our targets. In Figure 2B, 3'UTR ACADSB:c.*4177G>A is a perfect example of our significant hits, as it displays the clear distinction of the trends of the two genotypes across three fractions.

It is widely known that the alteration of polysome profiling distribution indicates the change of translational efficiency. Our GLM model helped us identify the UTR pairs whose WT and mutant have different polysome profiling patterns and thus likely have distinct translational efficiency. Nevertheless, since we only had limited polysome fractions in our experiments, we further validated our significant hits and confirmed the direction of regulation using luciferase reporter assay.

> *(3) The paragraph starting with "Even with the high confidence dataset, we did not intend to immediately translate the altered polysome profile into an increase or decrease in translation efficiency" is confusing. The whole premise of the screen used by the authors is that polysome profiling is a useful proxy for estimating levels of translation, so claiming that it doesn't necessarily measure translation is counterintuitive.*

In line with our reply to the last question, our goal is to use the alteration of polysome profiling patterns as a proxy for the change of translational efficiency. However, due to the limited number of fractions in our experiment, we could not directly infer the direction of regulation, i.e. increase or decrease of translational efficiency, of the statistically significant variants. That is why we refrained from making any conclusion about the direction of the regulation for the significant hits and proceed to validate them using luciferase reporter assay.

> *(4) Figure S5A - this is normalized to the nucleotide distribution in 5' or 3'UTRs? Is this statistic being applied to 27 SNPs in 3'UTRs?*

To identify sequence features associated with altered polysome association, we systematically analyzed both significant and nonsignificant UTRs for nucleotide and motif-level changes. Fisher's exact test was employed to evaluate whether specific nucleotide or motif alterations were enriched or depleted in polysome-shifting UTRs, compared to nonsignificant UTR pairs. For example, in the case of nucleotide C (see table below; also Table S4 and new Fig. S6A), only four significant 3' UTRs involved a change in C, resulting in a significant depletion of this nucleotide change among polysome-shifting 3' UTRs (odds ratio = 0.22, p = 0.0069). Expanding this approach to all 1-7 nt motifs, we identified multiple motif and nucleotide changes that were significantly associated with altered polysome association.

Author response table 2.

Nucleotide (C)	HC-Significant 3' UTRs	HC-Nonsignificant 3' UTRs
Change of nucleotide	4	94
No change of nucleotide	15	78

> *(5) "uATG in the 5' UTR was not identified by the model as a widespread feature explaining polysome shifting". Is this because of the method of ribosome profiling or because of the sequences in the library? Can having more sequences in the library specifically looking at 5'UTR give more power for such an effect to emerge?*

Our assay design accounted for the presence of upstream ATG codons and the strength of adjacent Kozak sequences. However, additional factors known to influence the function of upstream open reading frames (uORFs)—such as the reading frame of the uORF relative to

the main coding sequence, and the use of nonATG initiation codons—were not systematically included. As a result, the current assay may have limited sensitivity in detecting uORF-related regulatory effects. A dedicated design specifically tailored to uORF variants is likely to enhance the detection power and better capture their contribution to translational control.

> *(6) Figure 7B- it is not clear whether the luciferase reporter and the GFP reporter in the library function in a similar manner; is it creating out-of-frame or in of in frame uORF? Also, it is not clear if the differences are statistically significant.*

In the MPRA library, the IRF6 uORF is out of frame relative to the GFP coding sequence. To directly assess its translational impact, we employed a luciferase reporter assay by fusing luciferase downstream of the IRF6 uORF. These constructs revealed a significant reduction in protein production, as shown in Figures 3 and 7B–F. Although the clinically relevant IRF6 uORF is out-of-frame with the main ORF, we engineered an inframe uORF variant to validate translation initiation at the upstream ATG (uATG) (Fig. 7B-D). The in-frame construct confirmed uATG usage and led to a significant reduction in luciferase protein expression. Together, these results support the conclusion that the IRF6:c.-4609G>A variant gives rise to an active uORF that suppresses translation of the main ORF.

> ***Reviewer #2 (Recommendations For The Authors):***
>
> *(1) It would be helpful for the authors to subcategorize their data in ways that they consider meaningful and interpretable (e.g. shifts from all monosome to heavy, all heavy to monosome/free, etc.) Relatedly, what do the authors think the functional meaning is when a given transcript has high mono/heavy occupancy but low light occupancy (like what is shown in Figure 2B for ANK2) in the polysome profiling experiment? It is not apparent why a transcript with a high ribosome occupancy (heavy) would also have light occupancy (light).*

From the amplicon sequencing data, we obtained read counts for each UTR variant across the monosome, light, and heavy polysome fractions. Notably, this approach does not preserve the original relative abundance of transcripts among the three fractions. That is, despite a greater abundance of mRNAs in the heavy polysome fraction, comparable numbers of sequencing reads were recovered from the monosome and light fractions. As a result, this method is not suitable for interpreting the global directionality of translational shifts but is well-suited for detecting relative differences in polysome association. Therefore, our experimental and analytical design—combining targeted amplicon sequencing with generalized linear modeling (GLM)—was optimized to identify UTR variants that alter polysome association, independently of absolute transcript abundance in each fraction.

> *(2) The method put forward in Figure 2 would be more convincing if there was data showing reproducibility in the massively parallel reporter assay. Perhaps the mut/WT ratio for all transcripts can be plotted against each other and a statistical test of correlation can be performed.*

Thank you for pointing this out. To demonstrate the reproducibility of our massively parallel reporter assay, we have plotted scatter plots of the $ln(\frac{mut}{wt})$ ratios of all transcripts (summing the monosome, light, and heavy fractions) across different batches using our high-confidence dataset. We calculated the Pearson correlation coefficients and corresponding p-values for these comparisons. The results show strong correlation between each batch, supporting the reproducibility of our assay. We have incorporated this analysis in the main text as well as Supplemental Figure 3: “Pearson correlation analysis revealed R coefficients ranging from

0.59 to 0.71 for the mut-to-WT transcript ratios across three independent experiments (Supplemental Fig. 3)."

> *(3) The dots in Figure 2B indicate separate experiments, but the y-axis is log(counts). Values could be normalized (perhaps a ratio of mut/WT) for comparison between experiments.*

We aimed to compare UTR distribution across polysome fractions and recognized the importance of presenting the distribution patterns for both genotypes. This approach allows us to more clearly illustrate the differences or similarities in polysome association between the two genotypes.

> *(4) When describing the 5' UTRs used for the validation experiments in Figure 3, more information about the 5' UTR sequence used is necessary. It is not clear how much or what part of the 5' UTRs were removed, or why this was necessary considering the same experiment was conducted using full-length UTRs.*

In the initial library design, technical limitations of bulk oligonucleotide synthesis constrained the UTRs to 155 nucleotides, comprising 115-nt of endogenous human UTR sequence flanked by 20-nt priming sites on both ends. Variants were centered at the 58th nucleotide within the 115-nt UTR sequence. When one flanking region of the native UTR was shorter than 57 nt, the variant was shifted accordingly toward the shorter arm to maintain the 115-nt UTR length (Fig. 2A).

Given that endogenous UTRs in the human genome are often longer than 155 nt, we further evaluated the functional consequences of variants within full-length UTR sequences (Fig. 3B). While the mutant effects observed in the library setting were largely recapitulated, their magnitude was diminished in the full-length context, likely due to the increased sequence and structural complexity.

To clarify the experimental design related to Figure 3, we modified the text as the following: "The variants significantly altering the polysome profile were then individually validated by means of high-sensitivity luciferase reporter assays (Fig. 3A). To that end, we resynthesized both the variant and corresponding wildtype alleles in the same library format - 115-nt native UTR segments centered on the variant and flanked by 20-nt priming sites. These UTRs were then cloned upstream (5') or downstream (3') of the firefly luciferase coding sequence, depending on their genomic location." (p. 11)

> *(5) The conclusions from inserting RBP-binding motifs into 5' UTRs and assaying translational output (Figure 4) would be strengthened by including luciferase reporters containing endogenous 5' UTRs containing these motifs, and versions where the motifs are disrupted.*

Several variants that altered translation efficiency were validated in their native sequence contexts, including 5' UTR variants in *DMD* and *NF1* that affect SRSF1/2 binding sites, as well as a 3' UTR variant in *AL049650.1* that impacts a KHSRP binding site (Fig. 3 and Supplemental Figs. S1 & S2). To address the functional relevance of these variants within their native regulatory landscapes, we have incorporated the following clarification into the text (p. 13): "This observation is consistent with additional findings where variants that create or disrupt specific RBP binding sites—such as SRSF1/2 (e.g., in DMD and NF1; Fig. 2 and Supplementary Fig. S4) and KHSRP (e.g., in AL049650.1; Fig. 2 and Supplementary Figs. S4 & S5)—led to significant changes in translation efficiency within their native UTR contexts."

> *(6) Figure 5C shows that 5' UTR SNPs that form an uAUG are associated with greater structural changes, but this does not "indicate" that "structure-modifying UTR variants*

> *may control primary ORF translation partly by interfering with translation initiation from a uORF." The data presented in Figure 5 and luciferase/polysome data presented previously do not distinguish whether translation is occurring at an uAUG or canonical AUG. The statement quoted above is speculative and it should be clear that it is a hypothesis generated by the data and is not conclusive.*

We appreciate the reviewer's suggestion. We have therefore modified our text to: "Therefore, while changes in uATG may not be common explanatory factors for polysome-shifting mutations, our results suggest that structure-modifying UTR variants may control primary ORF translation partly by interfering with translation initiation from a uORF." (p. 14)

> *Minor points/questions*
>
> *(1) The authors should clarify whether during library construction for massively parallel polysome profiling the 3' UTR constructs contain a common 5' UTR? Likewise, do the 5' UTR constructs contain a common 3' UTR? Perhaps the lack of a 5' UTR in the 3' UTR constructs, which is implied by Figure 2A, would influence differences seen between 3' UTR pairs (and likewise for 5' UTR pairs).*

There are short common 5' UTRs appended to the 3' UTR library, and likewise, a common short 3' UTR is included in the 5' UTR library. The common 5' UTR comprises partial sequences from the CMV promoter and the plasmid backbone of pEGFP-N1 vector. The common 3' UTR includes sequences from the pEGFP-N1 backbone and a short polyadenylation signal from *HBA1* (hemoglobin subunit alpha 1). While we cannot entirely rule out potential crosstalk between 5' and 3' UTRs, the design ensures that all constructs are compared in a controlled and consistent context, enabling valid pairwise comparisons between variant and wildtype alleles.

To clarify the library design, we have revised the main text to include this explanation:

"The entire library of UTR oligonucleotides (UTR library) was subsequently ligated upstream or downstream of an enhanced GFP (EGFP) coding region, along with a CMV promoter and a common UTR sequence on the opposite end. Cells transfected with the UTR library were treated with cycloheximide 14 hours post transfection and then subjected to polysome fractionation (see Methods)." (p.11)

"The variants significantly altering the polysome profile were then individually validated through highsensitivity luciferase reporter assays (Fig. 3A). To this end, we resynthesized both the variant and corresponding wildtype alleles in the same library format - 115-nt native UTR segments centered on the variant and flanked by 20-nt priming sites. These UTRs were then cloned upstream (5') or downstream (3') of the firefly luciferase coding sequence, depending on their genomic location. As the initial library design, the test UTR segment differs only by one nucleotide, while a shared short UTR fragment is present on the opposite end of the coding sequence to ensure consistency across constructs (Fig. 2A)." (p. 12)

> *(2) The lines connecting the polysome distribution points make the plots appear busy and difficult to read, the data would be easier to interpret if they were removed.*

We employed a generalized linear model (GLM) to identify the variants that altered the polysome association of the corresponding transcripts. Statistically speaking, we were looking for the variants which led to significant interaction between genotype and polysome fractions. Ergo, displaying the lines as it is in our plots offers readers a convincing visualization of the interaction: lines from WT and Mut groups were not parallel, which indicates the interaction between genotype and polysome fractions. Moreover, showing the lines from three batches of experiments also helps us ascertain the reproducibility of our

experiments. Taken all together, the presence of the lines makes our plots even more informative.

https://doi.org/10.7554/eLife.98814.2.sa0